UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE QUÍMICA

Luciano Borges Censoni

Modelos baseados em contatos no estudo do enovelamento e da difusão térmica em proteínas

Campinas
2019

Luciano Borges Censoni

Modelos baseados em contatos no estudo do enovelamento e da difusão térmica em proteínas

Tese de Doutorado apresentada ao Instituto
de Química da Universidade Estadual de Campinas
como parte dos requisitos exigidos para a obtenção
do título de Doutor em Ciências.

Orientador: Prof. Dr. Leandro Martínez

O arquivo digital corresponde à versão final da Tese defendida pelo aluno Luciano Borges Censoni e orientada pelo Prof. Dr. Leandro Martínez.

Campinas
2019

Agência(s) de fomento e nº(s) de processo(s): CAPES, 001; FAPESP, 2013/08293-7
ORCID: https://orcid.org/0000-0002-2786-7910

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Química
Camila Barleta Fullin - CRB 8462

C332m

Censoni, Luciano Borges, 1988-
Modelos baseados em contatos no estudo do enovelamento e da difusão térmica em proteínas / Luciano Borges Censoni. – Campinas, SP : [s.n.], 2019.

Orientador: Leandro Martínez.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Química.

1. Redes complexas. 2. Termodifusão. 3. Enovelamento de proteínas. 4. Teoria da informação. I. Martínez, Leandro, 1979-. II. Universidade Estadual de Campinas. Instituto de Química. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Contact-based models in the study of folding and thermal diffusion in proteins
Palavras-chave em inglês:
Complex networks
Thermal diffusion
Protein folding
Information theory
Área de concentração: Físico-Química
Titulação: Doutor em Ciências
Banca examinadora:
Leandro Martínez [Orientador]
Adalberto Bono Maurizio Sacchi Bassi
Carlile Campos Lavor
Lucas Bleicher
Leonardo Paulo Maia
Data de defesa: 18-01-2019
Programa de Pós-Graduação: Química

BANCA EXAMINADORA

Prof. Dr. Leandro Martínez (Orientador)

Prof. Dr. Adalberto Bono Maurizio Sacchi Bassi (IQ/UNICAMP)

Prof. Dr. Carlile Campos Lavor (IMECC / UNICAMP)

Prof. Dr. Lucas Bleicher (Universidade Federal de Minas Gerais)

Prof. Dr. Leonardo Paulo Maia (Universidade de São Paulo - São Carlos)

A Ata da defesa assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Este exemplar corresponde à redação final da Tese de Doutorado defendida pelo(a) aluno(a) **LUCIANO BORGES CENSONI,** aprovada pela Comissão Julgadora em 18 de Janeiro de 2019.

"A sun of rubber was convulsed and set;
And blood-black nothingness began to spin
A system of cells interlinked within
Cells interlinked within cells interlinked
Within one stem. And dreadfully distinct
Against the dark, a tall white fountain played.

I realized, of course, that it was made
Not of our atoms; that the sense behind
The scene was not our sense. ...
. . . . . .
... But in the case
Of my white fountain what it did replace
Perceptually was something that, I felt,
Could be grasped only by whoever dwelt
In the strange world where I was a mere stray."

John Francis Shade, "Pale Fire: A Poem in Four Cantos",
em Vladimir Nabokov, "Pale Fire"

Agradecimentos

No momento em que escrevo estes agradecimentos, chega ao fim uma série de processos independentes de médio e longo prazo que conspiraram para terminar precisamente ao mesmo tempo. Nos últimos trinta dias, sem ordem específica de importância: comemorei meu trigésimo aniversário, me casei, finalizei a escrita desta tese e solicitei uma permissão de residência num país em outro continente. A confluência destes eventos me rouba em parte o distanciamento necessário para recordar todos os que estiveram presentes e agradecê-los com a generosidade devida, então peço desculpas àqueles que porventura tenha deixado de mencionar.

Agradeço aos meus pais Marcia e Olyntho pelo afeto e apoio cruciais desde o início desta trajetória acadêmica.

Agradeço ao Leandro, meu orientador, pela paciência e amizade ao longo destes anos somadas à excelente competência profissional.

Agradeço à Mariana, minha esposa, pelo amor e companheirismo, e pela paciência neste ano difícil. Amo você.

Agradeço a todos os companheiros do grupo, em particular a Mariana, o Gabriel e o Ivan, pelas discussões e pela companhia, e também ao Adriano pela disposição em ajudar.

Agradeço a todos os amigos que acompanharam o caminho de perto e de longe, Rodolfo e Milena, Pedro, Válter, Bruna, Lorrana, Rafaela, Caio, Sato, Laiz, e aos novos amigos que ganhei em 2016, Lucas, Jorge, Paulo, Akira, Thomás, Marcelo, César, Fábio, Henrique, a turma toda. Obrigado a todos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), projetos 2010/16947-9, 2013/05475-7 e 2013/08293-7.

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), projeto 470374/2013-6.

Resumo

Cadeias proteicas exibem uma escala natural de discretização que corresponde ao nível dos resíduos de aminoácido individuais; mecanismos evolucionários como inserção, deleção, ou mutações locais são incapazes de influenciar a identidade de um único átomo individual numa cadeia, e a totalidade da diversidade de estruturas e funções de proteínas se origina de combinações dos mesmos aminoácidos padrão em números e ordens variados. Aqui, modelamos estruturas de proteínas como redes de aminoácidos interagentes, e mostramos que características da topologia das redes resultantes podem ser usadas para predizer propriedades físico-químicas como a taxa da reação de enovelamento ou a capacidade de dissipar energia vibracional. Na primeira investigação, empregamos simulações modificadas de dinâmica molecular para medir a intensidade do acoplamento térmico entre cada resíduo e o resto da estrutura, e mostramos que medidas de importância de vértices derivadas da teoria de grafos, particularmente a centralidade de proximidade e a centralidade de autovetor, são bem correlacionadas com as intensidades dos acoplamentos observados. Construímos e resolvemos um modelo analítico de difusão térmica sobre uma rede, e mostramos que as melhores estimativas das intensidades dos acoplamentos térmicos resíduo-proteína podem ser derivadas da matriz Laplaciana que descreve a rede de interações. Na segunda investigação, ajustamos cadeias proteicas a um modelo baseado em caminhadas aleatórias auto-evitantes, e o empregamos para deduzir uma distribuição de probabilidades para a distância entre resíduos de aminoácidos em função de sua separação tomada ao longo da sequência. Usando esta distribuição, definimos uma expressão para a quantidade de informação probabilística associada à posição relativa de cada par de resíduos em uma estrutura. Mostramos que a quantidade de informação média sobre todos os pares de resíduos na estrutura tem boa correlação com o logaritmo de sua taxa de enovelamento. Subsequentemente, exploramos a mesma medida de informação para identificar contatos redundantes, e mostramos que é possível predizer a taxa de enovelamento de uma proteína com significativa precisão levando em conta menos de 5% de seus contatos. Finalmente, implementamos uma rotina para calcular ensembles de estruturas de proteínas sujeitas a restrições geométricas derivadas de experimentos de Ressonância Magnética Nuclear, e mostramos que a aplicação de um método de otimização baseado na estratégia da Otimização do Menor Valor Ordenado pode ajudar a distinguir as restrições que correspondem à atribuição correta de ressonâncias experimentais a partir de

conjuntos mistos que contém restrições corretas e incorretas.

Abstract

Protein chains exhibit a natural scale of discretization at the level of individual amino acid residues; evolutionary mechanisms such as insertion, deletion, or local mutation are unable to affect the identity of a single individual atom in the chain, and the full diversity of protein structures and functions stems from combinations of the same standard amino acids in various chain lengths and orders. Here, we model protein structures as networks of interacting residues, and show that features of the topology of the resulting networks can be used to predict physicochemical properties such as the rate of folding and the ability to dissipate vibrational energy. In the first investigation, we employ modified molecular dynamics simulations to measure the intensity of the thermal coupling between each residue and the rest of the structure, and then show that node importance measures derived from network theory, particularly closeness centrality and eigenvector centrality, are well correlated to the observed coupling strengths. We construct an analytically solvable model of heat diffusion on a network, and show that the best estimates of the intensities of residue-protein thermal couplings can be derived from the Laplacian matrix that describes the interaction network. In the second investigation, we fit protein chains to a model based on self-avoiding random walks, and use it to derive a probability distribution for the distance between amino acid residues as a function of their separation along the sequence. Using this distribution, we define an expression for the probabilistic information content associated to the relative position of each pair of residues in a protein structure. We then show that the average information content of all residue pairs in a structure is well correlated to the logarithm of its folding rate. Subsequently, we exploit the same measure of information content to identify redundant contacts, and show that we are able to predict a structure's folding rate to good accuracy while taking into account less than 5% of its contacts. Finally, we implement a routine to calculate protein structural ensembles subject to geometric restraints derived from Nuclear Magnetic Resonance experiments, and show that the application of an optimization method based on the Low Order-Value Optimization strategy can help distinguish the restraints that correspond to the correct assignment of experimental resonances among decoys.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ACPro	Banco de Dados de Cinética de Enovelamento de Proteínas do Amherst College (Amherst College Protein Folding Kinetics Database)
ATD	Difusão Térmica Anisotrópica (Anisotropic Thermal Diffusion)
DC	[Modelo de] Difusão-Colisão
DNA	Ácido desoxirribonucleico
LOVO	Otimização do Menor Valor Ordenado (Low Order-Value Optimization)
NC	[Modelo de] Nucleação-Condensação
NOE	Efeito Overhauser Nuclear (Nuclear Overhauser Effect)
PDB	Protein Data Bank
RMN	Ressonância Magnética Nuclear
RMSD	Desvio Quadrático Médio (Root Mean Square Deviation)
SARW	Caminhada Aleatória Auto-evitante (Self-avoiding Random Walk)
TM-Score	Template Modeling Score
TOCSY	Espectroscopia de Correlação Total (Total Correlation Spectroscopy)
wwPDB	Worldwide Protein Data Bank

Sumário

Capítulo 1

Introdução

Proteínas respondem por uma fração da ordem de 50% da massa seca de células vivas, podendo corresponder a um número total de um milhão a dez bilhões de moléculas por célula, a depender da complexidade do organismo[12]. Ademais, proteínas são responsáveis por, ou estão envolvidas em, essencialmente todos os processos químicos observados em organismos vivos, possibilitando comportamentos tipicamente associados à vida tais como autorreplicação, extração e emprego de energia proveniente do ambiente, detecção de estímulos e reação aos arredores[13].

Reconhecidas desde o século XVIII em função da faculdade de coagular quando expostas a ácidos ou altas temperaturas[14], as primeiras descrições da atividade catalítica de proteínas datam de meados do século XIX e início do século XX, quando experimentos revelaram a existência de substâncias, batizadas "enzimas", capazes de catalisar reações bioquímicas mesmo na ausência de células vivas[15]. Seu isolamento, cristalização e classificação geral como proteínas se deu na década de 1930[16]. A publicação das primeiras estruturas tridimensionais para mioglobina, hemoglobina e lisozima, por volta da década de 1960, colocou em foco a marcante interdependência entre atividade e estrutura em proteínas[17]. Desde então, estruturas atômicas tridimensionais têm sido importante matéria-prima para a elucidação de processos bioquímicos, celulares e fisiológicos, e envolvidas diretamente em trabalhos premiados com mais de um prêmio Nobel por década em média[18].

Na década de 1970, o Protein Data Bank (PDB) foi estabelecido para atuar como repositório central de estruturas resolvidas, buscando fornecer subsídios para a investigação do enovelamento de proteínas e complexos, do mecanismo catalítico de enzimas e do funcionamento de proteínas específicas[19]. Nas décadas subsequentes, o termo "cristalografia de proteínas" foi suplantado pelo mais geral "biologia estrutural", e foram resolvidas questões da época bem como outras que sequer haviam sido propostas, tais como os mecanismos pelos quais proteínas reconhecem moléculas de DNA, ribossomos sintetizam cadeias polipeptídicas ou vírus infectam células, além da observação de importantes evidências moleculares de ancestralidade comum de espécies distintas[19].

No presente, aplicações práticas na indústria farmacêutica vão além de revelar interações proteína-ligante e incluem a busca por novos alvos, novos modos de ação, e o desenvolvimento de proteínas com atividade terapêutica[20]. O PDB, agora uma organização internacional de nome Worldwide Protein Data Bank (wwPDB)[4], conta hoje com aproximadamente 135.000 estruturas resolvidas, a maioria ($\sim 90\%$) das quais por meio de experimentos de difração de raios-X, e o restante principalmente por experimentos de ressonância magnética nuclear (RMN) em solução[21].

1.1 Bioquímica de proteínas

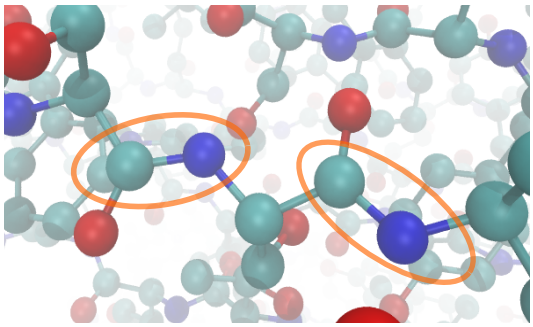
Quimicamente[1], proteínas são polímeros lineares de L-α-aminoácidos. Um aminoácido é um composto que consiste num carbono quaternário central, o carbono alfa ou C_α, ligado a um grupo carboxila, um grupo amina, um átomo de hidrogênio e um substituinte variável denominado *cadeia lateral* (vide figura 1.1). *In vivo*, proteínas são sintetizadas a partir de um conjunto limitado de tipos de aminoácidos, totalizando vinte comuns a todas as espécies mais um ou dois incorporados exclusivamente através de mecanismos traducionais raros e específicos[13]. Crucialmente, cada um dos tipos apresenta propriedades físico-químicas distintas. A variabilidade das cadeias laterais influencia desde volume e área de superfície a ponto isoelétrico ou a distribuição de doadores e aceptores de ligações de hidrogênio. A classificação mais natural e mais relevante é entre substituintes carregados, polares e hidrofóbicos.

Durante a síntese, cada novo aminoácido é incorporado à extremidade da cadeia nascente ligando-se covalentemente à carboxila exposta do anterior, ocasionando a expulsão de uma molécula de água. O fragmento remanescente, que passa a fazer parte da cadeia, é denominado resíduo de aminoácido, ou apenas *resíduo*. A ligação C-N entre resíduos vizinhos é denominada ligação peptídica, e tem caráter intermediário entre ligação simples e ligação dupla. Um exemplo é ilustrado na figura 1.1. A rotação em torno do eixo da ligação peptídica é fortemente inibida, resultando que os átomos ligados ao carbono e ao nitrogênio participantes residem todos no mesmo plano, e que os carbonos C_α de resíduos sucessivos adotam a conformação relativa *trans*[13]. Uma proteína consiste, deste modo, num encadeamento de carbonos C_α, unidos por ligações peptídicas planares e que expõem suas cadeias laterais em direções alternadas, delimitado pela amina do primeiro resíduo e a carboxila do último. Ao conjunto de todos os átomos à exceção das cadeias laterais dá-se o nome de cadeia principal ou *backbone* da proteína. Para muitas aplicações, a descrição da conformação espacial do backbone é suficiente, dispensando os detalhes das orientações das cadeias laterais. O comprimento de uma cadeia peptídica, medido em "número de resíduos", é frequentemente a dimensão relevante para fins de comparação entre estruturas proteicas. A massa total, reportada em kilodaltons (KDa),

[1] O texto desta seção inclui material adaptado de [3], do autor.

configura outra medida de interesse.

a)



b)

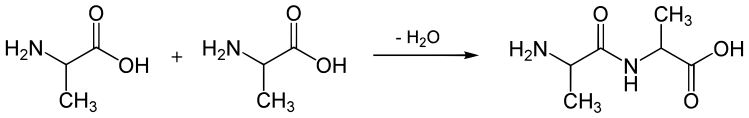


Figura 1.1. a) Destaque: exemplos de ligações peptídicas em uma estrutura proteica. Esferas vermelhas representam átomos de oxigênio, azuis representam nitrogênio e ciano representam carbono. Átomos de hidrogênio são omitidos. Figura gerada com o programa VMD[1, 2], e adaptada de [3]. b) Formação de ligação peptídica entre um par de aminoácidos, com perda de uma molécula de água. Ambos são Alaninas e têm CH_3 por cadeia lateral.

A sequência de aminoácidos ao longo da cadeia é denominada *estrutura primária*, e configura uma propriedade bem definida de cada proteína. Sob condições fisiológicas, o conjunto de conformações assumidas pela proteína em solução é univocamente determinado pela estrutura primária. Este conjunto ou *ensemble* de conformações, denominado *estado nativo* da proteína, é estável e reprodutível, em grande medida graças à estabilidade homeostática de condições como temperatura e pH no ambiente intracelular. A combinação das condições fisico-químicas locais com a natureza das interações intramoleculares e intermoleculares presentes em cada configuração atômica restringe o espaço conformacional que é termodinamicamente acessível para o ensemble como um todo, embora moléculas individuais possam ocupar momentaneamente configurações arbitrárias ao longo da evolução temporal do sistema. A conformação mais provável do conjunto é comumente denominada *conformação nativa*, introduzindo um pequeno abuso de notação. O espaço conformacional em si, bem como a sua fração acessível a uma dada cadeia, é de difícil representação visual em função do número considerável de graus de liberdade do sistema, ainda que desconsideradas as configurações do solvente. Contudo, postulando-se alguma noção de "localidade" tal que configurações suficientemente similares possam ser ditas *vizinhas* no espaço conformacional, é possível descrever características comuns à dinâmica de cadeias proteicas em geral.

Em solução, proteínas enoveladas são encontradas na maior parte do tempo em configurações que consistem em flutuações pequenas na vizinhança de uma ou poucas conformações de equilíbrio distintas. De fato, as conformações de equilíbrio ocupam o

fundo de “poços” ou “vales”, os mínimos da superfície de energia livre projetada sobre suas vizinhanças locais, e há autores que referem-se a cada região que compreende uma conformação de equilíbrio e sua vizinhança por “subestado”, terminologia que empregaremos ocasionalmente aqui[22, 23]. A existência de múltiplas conformações de equilíbrio é em geral associada a mecanismos de funções biológicas, das quais a catálise é um exemplo. Nestes casos, embora os mínimos alternativos possam ter energias livres comparáveis e, consequentemente, probabilidades de observação similares, as transições entre eles consistem em movimentos coletivos coordenados (ou por vezes difusivos e desorganizados) e requerem a passagem por estados intermediários de baixa probabilidade. Tais transições exibem, então, escalas de tempo da ordem de microssegundos a segundos, relativamente lentas do ponto de vista das escalas típicas de movimentos locais de átomos individuais, cadeias laterais ou *loops*. Flutuações associadas a movimentos rápidos e locais, por sua vez, são responsáveis pela variabilidade conformacional restrita a cada subestado[22, 23]. Ilustramos simbolicamente estes fenômenos na figura 1.2. As técnicas experimentais de estudos estruturais de proteínas diferem entre si também pela quantidade e tipo de informação que carregam sobre a variabilidade conformacional do ensemble como um todo ou sobre as flutuações em torno da conformação nativa.

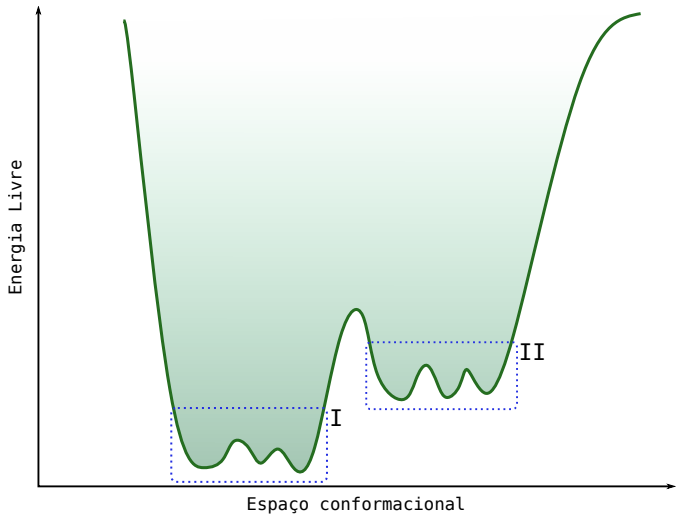


Figura 1.2. Ilustração representativa do espaço conformacional termodinamicamente acessível a uma cadeia proteica sob condições fisiológicas. Em representações desta natureza o alto número de graus de liberdade do sistema é condensado numa única coordenada conformacional sobre a qual se projeta a energia livre. Os mínimos I e II representam conformações de equilíbrio separadas por um estado de transição de baixa probabilidade, com taxa de interconversão da ordem de microssegundos a segundos. A rugosidade da superfície de energia livre reflete a variabilidade interna dos subestados, ocasionada por flutuações locais com escala de tempo muito mais rápida.

1.2 Estruturas atômicas tridimensionais

Observações experimentais de alta qualidade são capazes de delimitar as posições relativas dos átomos em uma conformação com precisões da ordem dos próprios diâmetros atômicos[24]. O conjunto das posições de todos os átomos de uma conformação, associado à informação previamente conhecida da topologia de suas ligações covalentes (derivada da estrutura primária de maneira previsível), é o que se reporta como estrutura tridimensional ou *estrutura* de uma proteína.

Nas estruturas determinadas por difração (ou *cristalografia*) de raios-X, a maioria das disponíveis, a conformação reportada é a mais provável (ou a média entre as conformações mais prováveis) sob condições de cristalização. Não obstante, o processo de cristalização frequentemente é suficientemente benigno para que a vizinhança imediata das moléculas de proteína não difira demais das condições de solução, incluindo quantidade importante de moléculas de solvente e íons associados ao soluto no cristal[25]. Deste modo, embora o ensemble do cristal seja mais restrito, a conformação mais provável é bastante similar à observada em solução. Diferenças importantes podem residir em regiões da macromolécula que apresentam maior flexibilidade no estado nativo; estas podem ser forçadas a assumir (ou aparentar) conformações muito mais rígidas que em solução[26].

Algumas propriedades do ensemble nativo podem ser recuperadas a partir da estrutura cristalina. O grau de variabilidade individual da posição de cada átomo é um subproduto do processamento algorítmico dos dados de difração de raios-X (especificamente, do refinamento da densidade eletrônica), e é reportado como um resultado experimental adicional, sob o nome de fator de temperatura ou B-fator (vide exemplos nas figuras 1.3 e 1.4). Além disso, propriedades dinâmicas podem ser recuperadas por meio de modelagem computacional a partir da estrutura estática. Em particular, características que dependem de dinâmicas majoritariamente harmônicas, tais como a variabilidade local do backbone, tendem a ser adequadamente reproduzidas a partir da análise de modos normais ou de modelos baseados em contatos intramoleculares. Conformações de cadeias laterais individuais, por outro lado, tendem a obedecer dinâmicas anarmônicas significativamente mais difíceis de modelar[27], assim como mudanças estruturais de grande escala.

A observação direta de conformações de solução, por outro lado, é possível mediante experimentos de RMN em proteínas[22]. Nestes, núcleos atômicos suscetíveis (principalmente ^{1}H, ^{13}C e ^{15}N) são levados a estados excitados de *spin* pela atuação combinada de campos magnéticos constantes e pulsos de radiofrequência. A frequência precisa da resposta nuclear depende do ambiente químico de sua vizinhança local; o espectro das respostas nucleares carrega consigo esta informação estrutural, permitindo em casos favoráveis associar a cada núcleo sua frequência ressonante. Protocolos experimentais mais sofisticados permitem então induzir a transferência de magnetização entre núcleos

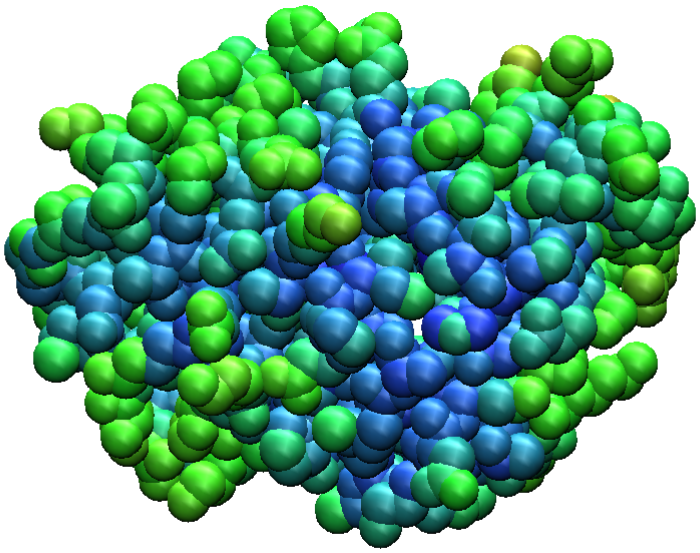


Figura 1.3. Representação da estrutura de uma hemoglobina (código PDB 1H97[4]) resolvida por difração de raios-X, com átomos colorizados em função do fator de temperatura experimental. Cores mais frias representam fator de temperatura menor, evidenciando que a mobilidade atômica tende a aumentar na direção do centro para a periferia da estrutura. A definição precisa do que é uma posição central será discutida no capítulo 3. Átomos são representados por esferas de raio igual ao raio de van der Waals correspondente. Figura gerada com o programa VMD[1, 2].

vizinhos, indiretamente através de ligações covalentes intermediárias ou por acoplamentos diretos através do espaço. Em ambos os casos, a dependência da intensidade dos acoplamentos com a distância ou a topologia, associadas ao conhecimento da identidade dos núcleos responsáveis pelas ressonâncias detectadas, permitem a dedução de *restrições geométricas* que vinculam pares de átomos distintos e possivelmente separados por muitos resíduos ao longo da cadeia[13]. Estas restrições geométricas, somadas à estrutura primária conhecida e a parâmetros físico-químicos como raios atômicos e energias potenciais de interação, dirigem o cálculo computacional de um conjunto de conformações compatíveis com as observações (este processo é descrito em maiores detalhes no capítulo 5). A variabilidade conformacional reportada neste conjunto muitas vezes é consequência de variações na densidade de restrições geométricas por resíduo, mas também reflete dinâmicas locais que não são necessariamente observáveis em experimentos de cristalografia[27]. A relação entre flexibilidade local e densidade de restrições é um fator importante que figura no algoritmo desenvolvido no capítulo 4 e será oportunamente discutida.

Em proteínas que dispõem de estruturas experimentais resolvidas tanto por cristalografia quanto por RMN, observa-se boa concordância entre os resultados, com similaridade média mais alta em aminoácidos hidrofóbicos e menos acessíveis ao solvente[28], de forma que as metodologias se complementam e validam mutuamente.

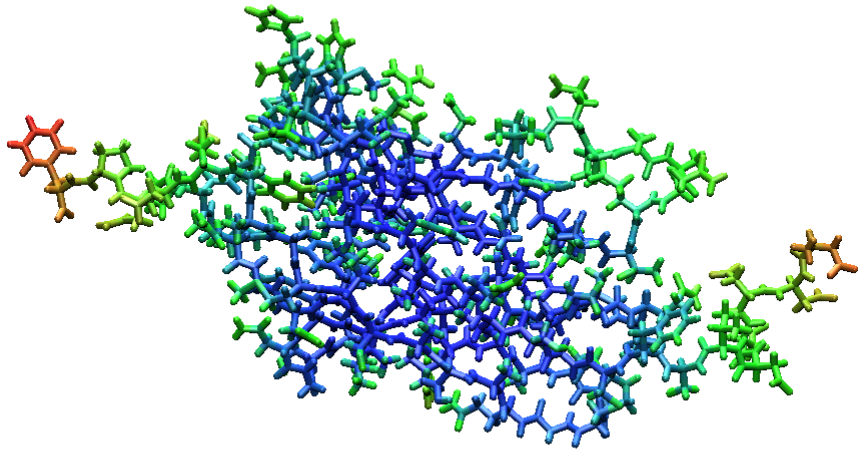


Figura 1.4. Representação da estrutura de um domínio imunoglobulina (código PDB 4CRP[4]) resolvida por RMN, com átomos colorizados em função do fator de temperatura reportado. Não existe definição única para o cálculo do fator de temperatura em experimentos de RMN, e os autores não reportam a metodologia utilizada. O mais provável é que a medida apresentada seja derivada da variabilidade na posição de cada átomo medida sobre o ensemble de modelos calculados. Cores mais frias representam fator de temperatura menor, evidenciando que a mobilidade atômica tende a aumentar na direção do centro para a periferia da estrutura, e se torna significativa nas extremidades da cadeia, além de ser mais alta em solução do que em condições de cristalização. Os tubos representam ligações covalentes com um átomo em cada extremidade. Figura gerada com o programa VMD[1, 2].

1.3 Medidas de similaridade estrutural

A flexibilidade intrínseca que define um conjunto de conformações relacionadas em torno de um dada estrutura impõe que não seja trivial decidir quando duas conformações quaisquer devem ser consideradas "a mesma", isto é, definir uma operação *identidade* sobre o espaço conformacional. A quantificação da similaridade entre duas estruturas tipicamente tem como pré-requisito a identificação das correspondências entre átomos, seguida da determinação dos movimentos de translação e rotação que levam à melhor sobreposição entre ambas. A esta transformação denomina-se *alinhar* as duas estruturas (vide figura 1.5), e "melhor" neste caso se define pela maximização da própria medida de similaridade, ou alternativamente a minimização de uma medida de desvio. A quantificação da similaridade se mostra, deste modo, um problema de otimização[29]. A função objetivo mais direta para tal aplicação é o desvio quadrático médio ou *root mean square deviation* (RMSD), reportada em Angstroms. Embora possam-se estabelecer limites superiores de RMSD para demarcar a relação de identidade entre estruturas, em geral da ordem de 2,0Å a 3,0Å (vide, por exemplo, [28]), a medida incorre numa desvantagem importante: não é independente do comprimento da cadeia[30]. Com isso, o mesmo valor de RMSD

pode indicar uma similaridade estrutural significativa ou inconsequente, a depender do par de estruturas comparadas.

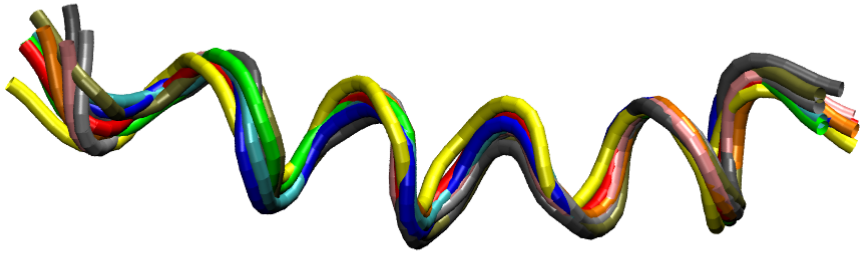


Figura 1.5. Exemplo de alinhamento estrutural de alta qualidade sobre dez modelos para um peptídeo pequeno (código PDB 2JQ0[4]), segundo critério de minimização de RMSD. O backbone de cada estrutura é representado por um tubo contínuo. Figura gerada com o programa VMD[1, 2].

Um exemplo de medida de similaridade mais sofisticada é o *template modeling score* ou TM-Score[31]. Definido no intervalo $(0, 1]$, o TM-Score implementa um formato funcional ligeiramente mais complicado (equação 1.1) para tornar-se efetivamente independente dos tamanhos das cadeias comparadas, desfrutando por isso de maior interpretabilidade.

$$\mathrm{TM} = \frac{1}{N}\left[\sum_{i=1}^{N_{\mathrm{equiv}}} \frac{1}{1 + d_i^2/d_0^2}\right], \quad d_0 = -1{,}8 + \sqrt[3]{N - 15} \tag{1.1}$$

Na equação 1.1, N é o número de resíduos da proteína alvo, N_{equiv} é o número de resíduos equivalentes entre as duas proteínas comparadas, determinado segundo o alinhamento estrutural que maxima o valor de TM, e d_i é a distância entre o i-ésimo par de resíduos equivalentes nesse alinhamento. A constante d_0 depende de N e é a normalização que torna a medida independente do comprimento das cadeias.

Via de regra, o TM-Score é menos sensível que o RMSD a diferenças locais em estruturas que são globalmente muito similares[30]. Demonstra-se que, em pares de estruturas comparadas sem alinhamento sequencial prévio, um TM-Score de 0,2 ou menor é equivalente à similaridade esperada entre proteínas escolhidas aleatoriamente, e, portanto, insignificante. Além disso, valores menores que 0,4 indicam que as estruturas não pertencem à mesma *classe* de enovelamento, enquanto valores a partir de 0,6 indicam que a classe é a mesma com alta probabilidade[30].

Propriedades como as citadas resultam que a medida de RMSD é por vezes preterida em aplicações reais em favor do TM-Score ou similares, em particular nas comunidades de modelagem molecular ou predição de estruturas de proteínas (vide, por exemplo, [32]). Todavia, cabe ressaltar que para aplicações desse caráter as medidas de

similaridade geral (*i. e.* não-local) são quase sempre calculadas levando-se em conta exclusivamente os carbonos C_α ou possivelmente só os átomos do backbone. Tal providência é necessária dada a variabilidade em geral maior das posições dos átomos das cadeias laterais, associada à tendência de medidas globais de similaridade terem seu valor dominado pelo maior desvio (tendência esta particularmente pronunciada para o RMSD).

Existe, ainda, outra classe de medidas de similaridade que não depende do alinhamento das estruturas a serem comparadas: medidas *baseadas em contatos*, que tabulam valores de distâncias internas ou de intensidades de interações na estrutura de interesse e comparam com as grandezas correspondentes em outra estrutura, contornando com isso a necessidade de resolver um problema de alinhamento global que admite múltiplas soluções[32]. Nestes casos, a própria definição do que é um contato pode ser consideravelmente simplificada ou *de baixa resolução* a depender da aplicação pretendida, minimizando a quantidade de átomos levados em consideração em cada resíduo. A existência e utilidade de medidas de similaridade baseadas em contatos é evidência de que propriedades estruturais e funcionais de proteínas são muitas vezes discerníveis a partir de um conjunto reduzido de parâmetros, e que descrições que incluem explicitamente todos os átomos podem ser escandalosamente redundantes. Ao longo do capítulo 5, empregamos com bons resultados a medida de TM-Score, em desfavor do RMSD ou de medidas mais complicadas. Contudo, a noção de descrições estruturais *reduzidas* de proteínas e suas aplicações é um tema que permeia este trabalho como um todo.

No capítulo 3, descrevemos estruturas de proteínas como modelos de redes de forma a evidenciar a topologia de seus contatos internos, e observamos que resíduos que ocupam posições privilegiadas na topologia se destacam também pela capacidade de difundir rapidamente sua energia térmica para o resto da estrutura.

No capítulo 4, desenvolvemos um modelo baseado em caminhadas aleatórias auto-evitantes para estimar a probabilidade de observação de contatos em cadeias enoveladas, e observamos importante relação entre as taxas experimentais de enovelamento e a verossimilhança de pequeno número de contatos importantes em cada estrutura.

No capítulo 5, investigamos a aplicação de um algoritmo de otimização que incorpora a identificação de *outliers* de energia ao problema do cálculo de estruturas de proteínas a partir de conjuntos de restrições geométricas que incluem medidas ambíguas ou incorretas.

Antes de mais nada, contudo, é necessário que apresentemos a metodologia utilizada para o cálculo da energia em estruturas arbitrárias, um passo crítico em quase todos os algoritmos que apresentaremos. Este será o tema do capítulo 2 a seguir.

Capítulo 2

Energia e campos de força em proteínas

Na seção 1.1, apontamos que as caraterísticas do espaço conformacional (ou conjunto de configurações) acessível para uma cadeia dependem das condições físico-químicas de seu ambiente local bem como de propriedades particulares de cada configuração, em particular de sua energia. Esta asserção é, na realidade, uma expressão resumida do fato de que proteínas, seja em condições experimentais de solução diluída ou no ambiente intracelular, são sistemas bem descritos pela termodinâmica estatística. Neste capítulo, oferecemos uma breve revisão dos conceitos associados, na medida da sua relevância para as investigações descritas.

2.1 Microestados e suas probabilidades

Uma proteína globular de duzentos resíduos, relativamente pequena, pode consistir de 3.000 átomos num caso típico. Um sistema constituído da mesma proteína imersa numa camada fina de solvente pode facilmente se aproximar de cinquenta mil átomos, com um número de graus de liberdade para posições três vezes maior. Naturalmente, a topologia das ligações covalentes restringe drasticamente o volume do espaço conformacional acessível ao sistema. Ainda assim, é evidente que o número total de configurações[2] diferentes que este pode apresentar é astronômico.

Aqui, cabe apontar que é habitual denotar uma configuração individual do sistema como um todo (*i. e.*, uma enumeração do valor assumido por cada grau de liberdade, ou por cada número quântico) por um *microestado*; um conjunto de microestados que respeitam alguma restrição macroscópica é denotado por um *macroestado*. Do mesmo modo, na discussão que segue o "espaço conformacional" sobre o qual o sistema transita representa, a rigor, o *espaço de fases*, visto que inclui também graus de liberdade para os momentos de cada partícula.

Dito isto, embora grande, o número de microestados acessíveis não cresce sem

[2]Embora os graus de liberdade sejam definidos sobre variáveis contínuas como posição ou velocidade, o fato de que no limite a dinâmica do sistema consiste de transições entre autoestados quantizados permite que continuemos essa discussão sem definir rigorosamente um procedimento de discretização que resulte em um "número" total de estados, assumindo apenas que tal procedimento é possível.

limites, pois todo sistema de interesse estará restrito aos macroestados compatíveis com as variáveis de estado locais. Para uma proteína em solução *in vitro*, tais variáveis podem ser a pressão atmosférica e a temperatura ambiente no laboratório, enquanto para a mesma proteína imersa no citoplasma parâmetros melhores podem ser a temperatura constante e o volume relativamente fixo da célula.

Já no caso de um modelo computacional atomístico da proteína como um sistema isolado, a própria energia total, constante, também representa um parâmetro macroscópico restritivo. Neste sistema em particular, a restrição da energia delimita precisamente o conjunto de microestados acessíveis, excluindo todas as configurações com energia total diferente daquela observada, mas nada diz sobre as probabilidades relativas de observação entre as configurações permitidas. De fato, desde que a dinâmica do sistema seja "suficientemente caótica", todos os microestados permitidos serão observados com a mesma probabilidade, ou, equivalentemente, o sistema dividirá seu tempo em média igualmente na ocupação de cada microestado.

Formalmente, a condição sobre a dinâmica do sistema é de que as trajetórias sobre o espaço de fases não apresentem *regularidades* (ou *periodicidades*) que impeçam sua exploração completa, tal que uma amostragem realizada sobre um tempo longo o suficiente seja equivalente àquela realizada sobre um conjunto descorrelacionado de condições iniciais[33]. Esta condição, denominada *ergodicidade*, é excessivamente difícil de provar analiticamente para quase qualquer sistema dinâmico à exceção dos mais simples[34], mas é na prática uma hipótese bem embasada por resultados experimentais em sistemas termodinâmicos.

A energia total constante é, contudo, privilégio de sistemas isolados. Para os sistemas que estudaremos aqui, a variável de estado relevante quase sempre será a temperatura constante, em geral fruto de troca energética mediante o contato com um sistema de dimensões muito maiores e temperatura bem definida que atua como *banho térmico*. Sob estas condições, a divisão entre microestados acessíveis e inacessíveis se torna muito menos abrupta, e as probabilidades de observação deixam de ser uniformes para variar continuamente com as energias totais dos microestados. Para um sistema sob temperatura termodinâmica T constante, a probabilidade p_i de observação de cada microestado é dada pela equação 2.1, que apresentamos aqui sem demonstração[35].

$$p_i = \frac{e^{-\frac{E_i}{kT}}}{\sum_i e^{-\frac{E_i}{kT}}} \tag{2.1}$$

Na equação 2.1, o índice i rotula microestados, E_i é a energia total do microestado i e k é a constante de Boltzmann.

Seja sob temperatura ou energia constantes, o tempo que o sistema permanece em cada macroestado será proporcional à soma das probabilidades do conjunto de microestados compatíveis. Sob energia constante, esta soma é simplesmente uma medida

do tamanho ou *volume* do macroestado no espaço de fases. De fato, fixada a energia, a cardinalidade deste conjunto é a única propriedade que diferencia macroestados em um mesmo sistema; a transição de um macroestado mais restritivo (um conjunto menor) para um menos restritivo (um conjunto maior) é sempre mais provável que a transição reversa. Para satisfazer propriedades desejáveis de aditividade em sistemas com múltiplos componentes, é natural trabalhar com uma transformação logarítmica desta medida, definindo, assim, a *entropia* do sistema (ou de macroestados específicos) sob energia constante como a equação 2.2.

$$S = k \ln \Omega \tag{2.2}$$

Na equação 2.2, Ω representa o número total de microestados acessíveis. Sob uma condição de temperatura constante, todavia, as probabilidades não são uniformes, e o número de microestados compatíveis não define isoladamente a probabilidade total. Assumindo a partir daqui por simplicidade também um volume fixo, esta passa a ser proporcional à soma na equação 2.3:

$$Z = \sum_i e^{-\frac{E_i}{kT}} \tag{2.3}$$

Quando a soma se estende por todos os microestados acessíveis ao sistema, ao invés de apenas um subconjunto referente a um macroestado específico, a equação 2.3 é precisamente o denominador da equação 2.1, atuando como constante de normalização das probabilidades. A soma Z é denominada a função de partição para este sistema, e Z e Ω cumprem papéis similares na medida em que a probabilidade de qualquer macroestado é dada pela razão entre sua função de partição (ou número de microestados) e a função de partição total (ou número total de microestados) para o sistema livre de restrições. Apesar disso, a relação de Z com a entropia não tem o mesmo formato que a relação de Ω (equação 2.2). A relação correta pode ser deduzida a partir da expressão estatística de Gibbs para a entropia de uma distribuição qualquer sobre microestados[36]:

$$S = -k \sum_i p_i \ln p_i \tag{2.4}$$

Note que ao substituirmos na equação 2.4 a probabilidade uniforme $p_i = 1/\Omega$ de cada um entre Ω microestados, recuperamos a expressão para a entropia sob energia constante da equação 2.2. Substituindo a probabilidade de Boltzmann (equação 2.1) ao invés disso, obtemos:

$$S = -k \sum_i \frac{e^{\frac{-E_i}{kT}}}{Z} \ln \frac{e^{\frac{-E_i}{kT}}}{Z}$$
$$S = -k \sum_i \frac{e^{\frac{-E_i}{kT}}}{Z} \left(\frac{-E_i}{kT} - \ln Z \right)$$
$$S = \frac{1}{T} \sum_i \frac{e^{\frac{-E_i}{kT}}}{Z} E_i + k \ln Z \frac{1}{Z} \sum_i e^{\frac{-E_i}{kT}}$$

Resultando:

$$S = k \ln Z + \frac{\langle E \rangle}{T} \tag{2.5}$$

A equação 2.5 é uma expressão para a entropia sob temperatura constante. Com formato comparável ao da equação 2.2, a diferença reside no termo proporcional a $\langle E \rangle$, que corresponde à média da energia *no ensemble*, isto é, a média da energia dos microestados ponderada por suas probabilidades[3]. É natural reorganizar a equação 2.5 para obter:

$$-kT \ln Z = \langle E \rangle - TS \tag{2.6}$$

A equação 2.6 define uma expressão para o logaritmo da função de partição Z em função de grandezas termodinâmicas macroscópicas como entropia, temperatura, e a energia interna do sistema, que assumimos corresponder à energia média calculada sobre o ensemble. Definimos previamente a função de partição correspondente a um dado macroestado como proporcional à soma das probabilidades dos seus microestados compatíveis (equação 2.3), e, portanto, uma medida da sua probabilidade total de observação. Agora, dispomos de uma relação que quantifica essa probabilidade em termos de grandezas macroscópicas; o macroestado de maior probabilidade será aquele que minimiza a soma $F = \langle E \rangle - TS$, denominada *energia de Helmholtz* mas frequentemente referida por "energia livre de Helmholtz" ou apenas "energia livre".

Na maioria das situações em que se pretende encontrar o macroestado que minimiza F, e certamente em se tratando da predição de estruturas de proteínas, é mais viável construir algoritmos que minimizem o termo de energia. Embora os dois termos cumpram papéis equivalentes, maximizar a entropia em geral não é um problema conceitual fácil de operacionalizar, e quase sempre é um problema computacional muito mais

[3] A equação 2.5 pode ainda ser colocada no formato $S = k \ln \left(Z / \exp \frac{-\langle E \rangle}{kT} \right)$ ou $S = k \ln \sum_i \exp \frac{\langle E \rangle - E_i}{kT}$, ressaltando sua similaridade em relação à equação 2.2 e sugerindo que esta, por sua vez, também pode ser pensada qualitativamente como $S = k \ln \sum_\Omega \exp \frac{\langle E \rangle - E_i}{kT} = k \ln \sum_\Omega 1$, pois todos os microestados têm a mesma energia. Assim, caracteriza-se a entropia como medida da *imprevisibilidade* da distribuição da energia sobre os microestados acessíveis, um tema que será revisitado no capítulo 4.

intratável. Assim, é natural propor algoritmos que procuram estruturas de alta probabilidade por meio da minimização da energia, sabendo que este processo guarda ao menos uma correlação parcial com a minimização da energia *livre*. Mais ainda, nos sistemas que estudaremos, sabemos que a energia total de cada microestado tem um formato geral como da equação 2.7, contendo em si termos relativos às posições e também aos momentos das partículas.

$$E_i(\vec{r_i}, \vec{v_i}) = V(\vec{r_i}) + K(\vec{v_i}) = V(\vec{r_i}) + \sum_{partículas} \frac{1}{2} m v_i^2 \tag{2.7}$$

Apontamos, porém, que a dependência da energia nos mesmos é aditiva, resultando que a probabilidade correspondente pode ser decomposta em termos multiplicativos separados para as energias potencial, função das posições, e cinética, função das velocidades. Para objetivos como os descritos neste trabalho, é praxe lançar mão desse expediente e trabalhar em termos da probabilidade que é função apenas da energia potencial e é maximizada pela minimização dos termos correspondentes, considerando que as velocidades não interferem nas probabilidades relativas das configurações e serão determinadas independentemente obedecendo a uma distribuição de Maxwell-Boltzmann. Tal procedimento será empregado frequentemente ao longo deste trabalho.

Voltaremos a discutir esta noção no capítulo 5, durante a apresentação da metodologia lá utilizada. Devemos ressaltar, contudo, que no intuito de simplificar esta apresentação, deixamos de mencionar alguns detalhes, tais como as correções necessárias para tratar de sistemas com partículas indistinguíveis, a troca da energia média pela *entalpia* na energia livre quando o sistema está sujeito a uma pressão constante ao invés de um volume fixo, ou o procedimento exato de discretização para transformar uma região do espaço de fases num número total de microestados. Estas correções, embora corretas e necessárias, não mudariam as conclusões apresentadas, e o leitor interessado é dirigido a um livro-texto como [35] para um tratamento completo.

Por ora, apresentaremos na seção 2.2 a metodologia empregada para o cálculo da energia potencial em proteínas em todos os trabalhos apresentados aqui.

2.2 Energia potencial em proteínas

Na seção 2.1, repassamos brevemente alguns conceitos que sugerem uma relação (ainda que parcial) entre a energia potencial e a probabilidade de observação para uma estrutura candidata qualquer. A exploração desta relação com o objetivo de produzir estruturas de alta probabilidade, que trataremos ocasionalmente também por estruturas de alta *qualidade*, somente é possível quando se pode expressar a energia potencial em função da configuração atômica, particularmente num formato compatível com a aplicação de rotinas de otimização. A viabilidade de tal metodologia depende, então, da disponibilidade de uma função energia potencial ao mesmo tempo computacionalmente amigável e

suficientemente precisa. A uma tal função dá-se o nome de *campo de força*.

É natural, nessas condições, restringir-se a uma descrição totalmente clássica da matéria, tratando cada átomo como uma carga pontual cuja posição e velocidade são conhecidas com precisão arbitrária. De fato, para os sistemas estudados, praticamente qualquer efeito que se deva aos elétrons ligados acontece numa escala de tempo muito mais rápida do que o movimento dos núcleos, e pode ser ignorado ou subsumido na forma de cargas parciais médias com segurança. Ademais, nas condições que se pretende modelar, não há efeitos significativos devidos a forças que dependam das velocidades, e a energia pode ser escrita como um potencial que é função apenas das posições atômicas. Assumir essas aproximações não restringe seriamente a variedade de formatos funcionais possíveis para a energia, e a literatura ainda disponibiliza opções concorrentes para o trabalho com estruturas de proteínas[37]. Escolhemos aqui implementar o campo de força CHARMM[5], que apresenta propriedades desejáveis de simplicidade e precisão por ser otimizado para simulações de dinâmica molecular. Seu formato geral é apresentado a seguir[4].

$$V = V_{\text{Coulomb}} + V_{\text{LJ}} + V_{\text{ligação}} + V_{\text{ângulo}} + V_{\text{diedro}} \tag{2.8}$$

Na equação 2.8, os dois primeiros termos referem-se à energia das interações entre átomos considerados não-ligados. Isso inclui todos os pares de átomos que pertencem a moléculas diferentes, bem como a maior parte dos pares de átomos pertencentes à mesma molécula, desde que separados por mais de três ligações covalentes consecutivas. Pares de átomos não-ligados interagem somente através de forças de Coulomb (equação 2.9), se possuírem carga elétrica, e de forças de van der Waals modeladas como um potencial de Lennard-Jones (equação 2.10).

$$V_{\text{Coulomb}} = \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{2.9}$$

$$V_{\text{LJ}} = \sum_i \sum_{j>i} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.10}$$

Em ambas as equações, as somas são sobre as listas de átomos interagentes. Na equação 2.10, as constantes σ_{ij} e ε_{ij} são obtidas a partir de combinações simples de constantes tabuladas para cada tipo de átomo individual[5]. Grandezas como estas são denominadas *parâmetros* do campo de força; uma vez fixado o formato funcional de cada termo de energia, respeitando um compromisso entre complexidade computacional e precisão, as constantes livres são ajustadas para que o campo de força reproduza o melhor possível resultados experimentais. Cargas e massas atômicas, constantes de força para ligações e outras interações são parametrizadas de forma que o campo possa reproduzir o melhor possível desde propriedades locais, como comprimentos de ligação, passando

[4]Nesta seção utilizamos material adaptado de [3], do autor.

por espectros vibracionais até propriedades termodinâmicas como a densidade e o ponto crítico do solvente.

Os três últimos termos da equação 2.8 referem-se às interações entre átomos ligados, isto é, conjuntos de átomos que participam diretamente de ligações covalentes, ou que através de ligações consecutivas definem ângulos ou diedros. Seus formatos são dados por:

$$V_{\text{ligação}} = \sum_{\text{ligação i}} k_i^{\text{ligação}}(r_i - r_{0i})^2 \tag{2.11}$$

$$V_{\text{ângulo}} = \sum_{\text{ângulo i}} k_i^{\text{ângulo}}(\theta_i - \theta_{0i})^2 + k_i^{\text{UB}}(S_i - S_{0i})^2 \tag{2.12}$$

$$V_{\text{diedro}} = \sum_{\text{diedro i}} \begin{cases} k_i^{\text{diedro}}[1 + \cos(n_i\phi_i - \gamma_i)], & n_i \neq 0 \\ k_i^{\text{diedro}}(\omega_i - \omega_{0i})^2, & n_i = 0 \end{cases} \tag{2.13}$$

Ligações covalentes (equação 2.11) são modeladas como um potencial harmônico simples em torno da distância de equilíbrio. Pares de átomos ligados covalentemente a um mesmo átomo central ficam sujeitos a um potencial harmônico em função do ângulo entre si (equação 2.12), em alguns (poucos) casos corrigido por um termo adicional dependente da distância denominado termo Urey-Bradley[5]. Conjuntos de quatro átomos unidos por ligações covalentes consecutivas definem um *ângulo de torção* ou *ângulo diedral* entre os planos que contém os três primeiros e os três últimos átomos, experimentando um termo de energia parametrizado como uma função periódica do ângulo de torção, com uma ou mais posições de equilíbrio. Já conjuntos de quatro átomos em uma cadeia ramificada definem *diedros impróprios*, modelados como funções harmônicas do ângulo entre o átomo central e o plano que contém os outros três. Ambos são contidos na equação 2.13. As três equações dependem, também, de parâmetros específicos que definem constantes de força e posições de equilíbrio. Ilustramos cada um destes termos na figura 2.1.

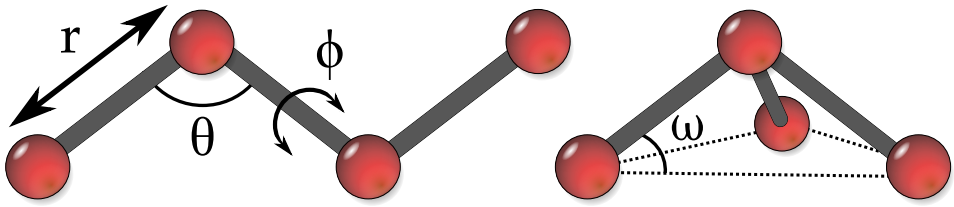


Figura 2.1. Ilustração dos termos de energia incidentes sobre átomos ligados do campo de força CHARMM[5]. À esquerda, a distância entre um par de átomos ligado covalentemente, o ângulo entre três átomos consecutivos e o ângulo diedral entre dois planos de três átomos. À direita, o diedro impróprio definido por uma cadeia ramificada.

Tomados em conjunto, os termos da equação 2.8 representam uma aproximação muito boa e de cálculo relativamente simples da energia interna para conformações ar-

bitrárias de estruturas proteicas. A equação 2.8 será, por isso, extensamente empregada no capítulo 5, no contexto das probabilidades discutidas na seção anterior. Mas a existência de uma expressão para a energia interna possibilita também uma exploração computacional mais direta do espaço configuracional: o cálculo da evolução dinâmica do sistema através da solução numérica das equações de movimento, uma estratégia denominada *dinâmica molecular*.

Embora neste trabalho não tenhamos realizado simulações de dinâmica molecular diretamente, dados provenientes de simulações adaptadas são parte importante das análises discutidas no capítulo 3. Por esta razão, repassamos brevemente conceitos da metodologia na próxima seção.

2.3 Simulações de Dinâmica Molecular

Na seção 2.1, observamos como o tempo médio que o sistema passa ocupando um estado macroscópico dado depende das probabilidades de observação do conjunto de estados microscópicos compatíveis com aquele. Embora determinar a probabilidade de observação de um único microestado seja simples, bastando para isto conhecer sua energia, no caso geral não existem bons procedimentos para enumerar todos os microestados compatíveis com um macroestado. Na prática, é necessário calcular probabilidades aproximadas para cada macroestado, baseadas em algum tipo de amostragem dos microestados compatíveis e o tempo em que o sistema os ocupa. Dentre vários protocolos possíveis para este tipo de *amostragem termodinâmica*, uma ideia relativamente direta é simular a evolução temporal do sistema partindo de uma condição inicial conhecida, e coletar as probabilidades a partir da trajetória dinâmica do sistema.

Aqui, continuamos sob as hipóteses adotadas na seção anterior; tratamos as estruturas de proteínas e quaisquer outras moléculas, tais como ligantes, íons ou solvente, como conjuntos de átomos com posições e velocidades determinadas com precisão arbitrária, sob a ação de forças determinadas por equações Newtonianas simples. De fato, sob uma descrição totalmente clássica, as forças decorrem imediatamente da energia potencial:

$$m_i \ddot{\vec{r}}_i = -\nabla_i V(\vec{r_1}, \vec{r_2}, \ldots, \vec{r_N}) \tag{2.14}$$

Naturalmente, a energia na equação 2.14 é o próprio campo de força descrito na seção anterior. Em primeira aproximação, observar a evolução das configurações atômicas sob essas equações de movimento seria equivalente a acompanhar a evolução do sistema em tempo real com um aparato experimental de altíssima precisão, e de fato simulações de dinâmica molecular oferecem muitas vezes a única alternativa para observar determinados mecanismos moleculares. A solução analítica destas equações é, contudo, evidentemente inviável para qualquer sistema de interesse, dados os múltiplos acoplamen-

tos entre as equações de movimento de cada coordenada. Ainda assim, existem estratégias de integração que permitem propagar soluções numéricas por muitos passos, mantendo características de performance e estabilidade aceitáveis. Talvez surpreendentemente, a estratégia mais direta *não é* uma delas. Calcular as posições e velocidades aproximadas a um tempo $t + \Delta t$ mediante uma expansão de Taylor de primeira ordem em torno de $\vec{r}(t)$ ou $\vec{v}(t)$, denominado método de Euler, incorre em sérios problemas de estabilidade para tempos longos e acumula um erro global de ordem linear em Δt[38]. Métodos de integração mais sofisticados são necessários, e o empregado aqui e na grande maioria das aplicações em dinâmica molecular é o algoritmo Verlet-velocidade, com melhor estabilidade e erro global da ordem de $(\Delta t)^2$ sem grande aumento do custo computacional[39]. A dependência dos erros acumulados com Δt põe em evidência também o compromisso entre precisão e custo computacional, na medida em que diminuir o erro esperado requer a subdivisão do tempo real simulado em um número maior de passos calculados.

De todo modo, pressuposta a disponibilidade de tempo computacional, o cálculo de trajetórias longas permite a observação direta dos tempos de residência ou número de visitas do sistema ao estado de interesse, e com isto a caracterização de suas propriedades termodinâmicas e a comparação com medidas experimentais. Para permitir comparações em equivalência de condições, é frequentemente preferível simular o sistema sob temperatura constante, uma sofisticação em relação à energia total constante que já é assegurada pela integração correta das equações de movimento.

As técnicas utilizadas para simular o acomplamento do sistema a um banho térmico são denominadas "termostatos"[40]. Aqui, o método utilizado é a integração da equação de movimento original perturbada pela adição de um termo dissipativo e um estocástico, denominada equação de Langevin[39]. Os termos adicionais atuam em conjunto para corrigir a energia cinética média na direção do valor desejado. Algoritmos similares existem também para o controle da pressão.

A combinação do algoritmo de integração, do termostato e de um campo de força adequado é suficiente para realizar o que se entenderia por uma simulação de dinâmica molecular padrão, amostrando configurações atômicas de acordo com suas expectativas termodinâmicas. Em alguns casos, contudo, modificações podem ser incorporadas para acelerar a amostragem ou aproximar o sistema simulado de situações experimentais específicas. Aqui, utilizamos dados provenientes de simulações adaptadas para representar a dissipação de energia a partir de uma condição inicial correspondente a uma perturbação térmica localizada. O protocolo utilizado nestas simulações é denominado Anisotropic Thermal Diffusion (ATD)[41, 42].

No protocolo ATD, estabelece-se um gradiente local de temperatura mediante o acoplamento seletivo de apenas um setor do sistema ao banho térmico. O gradiente é mantido por um tempo curto, durante o qual o fluxo de energia é observado e registrado. Detalhadamente, um experimento de ATD em uma estrutura de proteína consiste nos

seguintes passos:

(i) Um conjunto de conformações inicias aleatórias são amostradas a partir de uma distribuição de equilíbrio a 300K, e cada uma é subsequentemente re-equilibrada a uma temperatura de 10K.

(ii) Para cada conformação resfriada, um resíduo de aminoácido é selecionado e seus átomos são acoplados ao banho térmico quente (300K), enquanto o banho térmico frio é desligado. O sistema é simulado por um tempo curto fixo da ordem de picos-segundos, durante os quais as temperaturas da proteína e de cada resíduo individual são periodicamente registradas.

(iii) O passo anterior é repetido para cada resíduo de cada conformação inicial. Após todas as simulações, e supondo que os resultados tenham convergido adequadamente, os resultados são tipicamente apresentados na forma de gráficos de temperatura final de cada resíduo, ou da proteína como um todo, em função do resíduo aquecido.

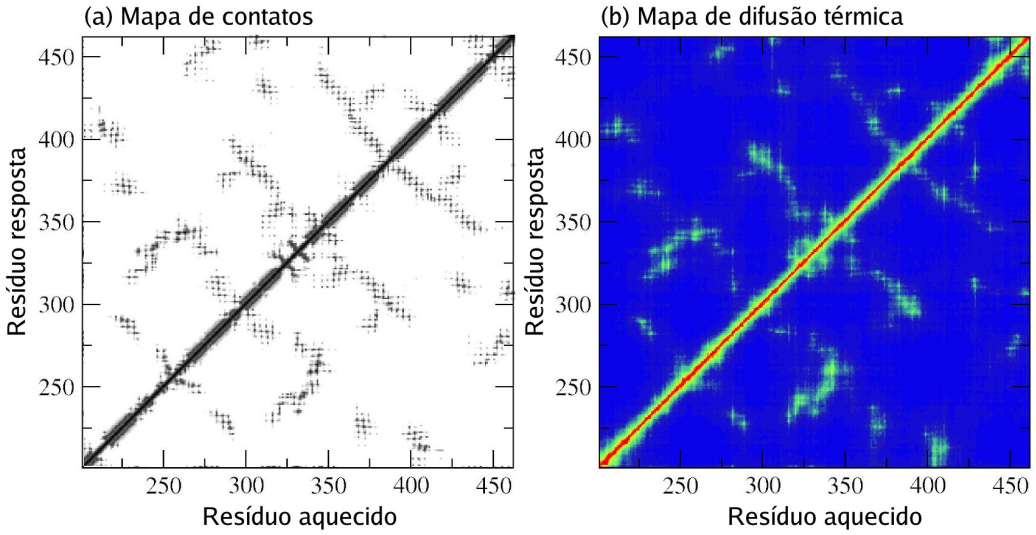


Figura 2.2. Resultados de um experimento de ATD, reportados na forma de um mapa de difusão térmica (direita), tabulando a temperatura final atingida por cada resíduo em função do resíduo aquecido. À esquerda, o mapa de contatos para a mesma proteína; a similaridade entre ambos evidencia a forte relação entre a topologia das ligações e a capacidade de difundir calor. Reproduzido com autorização de [6].

O protocolo descrito mimetiza uma condição experimental em que uma molécula de proteína é sujeitada a uma excitação externa que incide sobre um resíduo específico, tal qual a absorção de radiação ou a catálise de uma reação química[42]. Tendo em vista que o processo observado ocorre fora do equilíbrio, impõe-se que as simulações sejam realizadas com a mesma duração fixa para que os resultados sejam comparáveis. Ainda

assim, a temperatura final atingida pela proteína varia substancialmente em função de qual resíduo se acopla ao banho térmico quente. Tal resultado é particularmente interessante na medida em que resíduos identificados como especialmente capazes de dissipar calor sobre o resto da estrutura frequentemente coincidem com aqueles que destroem a atividade quando substituídos em experimentos de mutagênese sítio-dirigida[42].

Embora possa-se conjecturar que a variação da capacidade de difundir calor dependa da massa do resíduo aquecido, de seu caráter físico-químico ou sua acessibilidade ao solvente, a inspeção dos resultados sugere que o parâmetro dominante em primeira ordem é o tipo e número de interações das quais o mesmo participa. Um exemplo é ilustrado na figura 2.2. O capítulo 3, a seguir, é dedicado à formalização desta observação; construímos e resolvemos analiticamente um modelo simples que reproduz adequadamente os resultados obtidos por simulações de ATD, destacando a importância da topologia na difusão de energia em proteínas.

Capítulo 3

Modelos de rede para difusão térmica em proteínas

Neste capítulo, descrevemos investigações realizadas com o objetivo de expandir e complementar os resultados obtidos durante o projeto de mestrado do autor[3], que tratou da modelagem da proteínas como redes de aminoácidos interagentes e a subsequente aplicação de ferramentas de Teoria de Redes para a identificação de resíduos importantes no contexto da difusão térmica. Apresentamos aqui, de forma resumida, a motivação por detrás de tal investigação, bem como uma breve revisão dos resultados apresentados por ocasião do fim daquele projeto, com o objetivo de contextualizar os experimentos realizados posteriormente. Em seguida, apresentamos os desenvolvimentos que levaram à forma final do trabalho, publicado em [43].

3.1 Difusão térmica em proteínas

Apresentamos no capítulo 1 algumas características físico-químicas gerais de proteínas, enfatizando seu caráter dinâmico descrito como um ensemble de múltiplas conformações. No capítulo 2, discutimos brevemente as leis que regem as probabilidades de ocupação de cada conformação, e como estas probabilidades podem ser calculadas por amostragem em experimentos de simulação computacional. Até aquele momento, estivemos restritos ao caso particular em que a estrutura estudada está em equilíbrio térmico com sua vizinhança, com temperatura bem definida e, consequentemente, probabilidades relativas fixas. Apesar disso, gradientes ou variações temporais de temperatura podem revelar em proteínas mecanismos de resposta ou adaptação que não são necessariamente previsíveis a partir de seu comportamento de equilíbrio.

Um exemplo pode ser observado quando uma proteína é submetida a uma perturbação localizada, que leva um determinado resíduo a um estado vibracional consistente com uma temperatura mais alta que a do resto da estrutura. A relaxação de um tal estado excitado se dá, tipicamente, mediante transições entre estados vibracionais caracterizados por deslocamentos progressivamente menores de um conjunto de átomos progressivamente maior, à medida que a estrutura retorna ao equilíbrio. A transferência de energia entre estados vibracionais distintos é possibilitada por acoplamentos que dependem do próprio

conjunto de interações, covalentes e não-covalentes, entre átomos ou resíduos vizinhos. A topologia das interações, contudo, impõe que os efeitos de perturbações desta natureza não se propaguem isotropicamente sobre a estrutura, definindo "canais preferenciais" através dos quais a energia se dissipa mais rapidamente[44–46]. A figura 3.1 ilustra esse conceito.

Quando não é dissipada eficientemente, uma perturbação dessa natureza pode precipitar o início de uma reação de desnaturação[47]. De fato, em algumas proteínas termofílicas e hipertermofílicas, a manutenção da atividade catalítica parece depender da capacidade de direcionar excessos de energia para regiões periféricas móveis, enquanto a estruturação e flexibilidade adequadas são mantidas nos sítios funcionais[48].

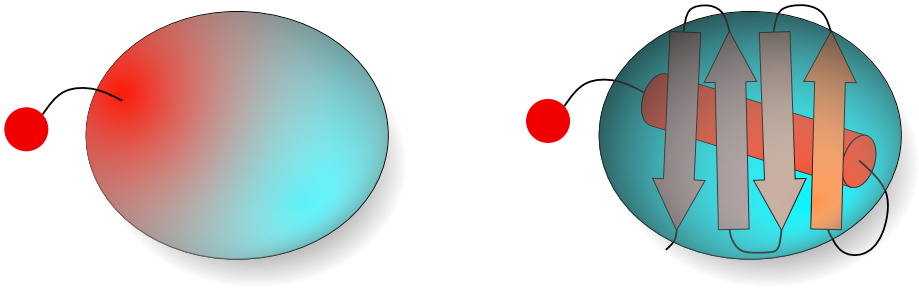


Figura 3.1. Ilustração do mecanismo referido no texto por "difusão anisotrópica" de calor. À esquerda, representação de um gradiente transiente de temperatura, induzido por uma perturbação local que, a partir do ponto de aplicação, se dissipa isotropicamente através de um meio uniforme. À direita, uma perturbação localizada incide sobre a extremidade de uma cadeia polipeptídica. O efeito das distribuições não-uniformes e heterogêneas de cavidades e de contatos no interior da estrutura pode resultar na propagação ao longo de canais privilegiados, de tal forma que regiões distantes do ponto de aplicação da perturbação podem atingir temperaturas mais altas que regiões adjacentes, se o acoplamento em relação às últimas for de menor intensidade.

O mesmo tipo de fenômeno se observa em mecanismos de alosterismo, quando a associação a um ligante ocasiona uma perturbação que se transmite e efetua uma mudança dinâmica ou conformacional em uma região distinta e possivelmente distante da estrutura[49–52]. Nestes casos, a existência de canais preferenciais pode atuar no sentido de retardar uma dissipação desorganizada da energia, permitindo a transmissão através do interior da estrutura até um sítio funcional secundário[53].

Independentemente das suas consequências funcionais, a existência de canais sobre os quais a propagação de energia é privilegiada parece ser uma propriedade intrínseca de estruturas de proteínas. Propriedades como a densidade e a compressibilidade de proteínas são similares às de sólidos, mas a distribuição espacial interna se assemelha a um empacotamento aleatório de esferas, com alta variabilidade nos raios das cavidades internas e densidade de empacotamento próxima do limiar de percolação[54]. Em aglomerados de percolação, canais de transporte surgem naturalmente como consequência de sua

geometria[55], sugerindo que no interior de proteínas a existência de canais preferenciais pode ser resultado do mesmo fenômeno. Esta observação é parte da motivação para os experimentos realizados neste capítulo.

Por seu caráter eminentemente dinâmico e transiente, a propagação de energia fora do equilíbrio como nos contextos citados é de difícil observação experimental. Embora existam alternativas como experimentos de relaxação de spin em RMN[56], neste trabalho nos restringimos a técnicas computacionais de modelagem, e na seção 2.3 apresentamos o protocolo ATD de simulações modificadas de dinâmica molecular. O protocolo ATD é uma técnica para acompanhar a propagação de energia térmica em proteínas, tabulando a resposta de cada resíduo individual a partir de um estado inicial que simula uma perturbação localizada. Os resultados de simulações de ATD tipicamente ressaltam o caráter anisotrópico e dependente da topologia da propagação da energia, na medida em que os mapas de difusão térmica observados guardam forte semelhança com o mapa de contatos entre resíduos na estrutura nativa[42].

Inserido nesse contexto, o projeto de mestrado do autor investigou o uso de redes complexas para modelar estruturas de proteínas enfatizando a topologia de suas ligações, e a subsequente aplicação de ferramentas de teoria de redes para identificar resíduos importantes para a difusão de energia. Na próxima seção, apresentaremos a metodologia empregada e os resultados obtidos.

3.2 Proteínas como redes de aminoácidos

Uma *rede* é uma representação de um conjunto de partes interagentes e suas interações. Especificamente, uma rede é uma abstração que enfatiza a presença ou ausência de relações entre participantes ao mesmo tempo em que despreza as propriedades individuais dos elementos e das interações. Assim, se torna uma ferramenta apropriada para descrever sistemas em que os efeitos das diferenças entre seus componentes e nas maneiras como eles interagem são desprezíveis em relação à *topologia* das interações[57].

A literatura oferece exemplos para uma ampla gama de sistemas que, sob algum regime de condições, se encaixam nessa definição. Por brevidade, não nos estenderemos neste preâmbulo fazendo uma enumeração de sistemas específicos ou de áreas do conhecimento para os quais o paradigma de redes tem sido aplicado com sucessos importantes; direcionamos o leitor interessado a [57], [58] ou [7]. Não obstante, a importância da ciência de redes está bem estabelecida, a ponto de que termos oriundos de seu léxico como "*hub*" ou "mundo pequeno" ultrapassaram os limites do seu próprio campo, e a metáfora da rede ou teia é hoje repetidamente invocada não apenas para descrever mas também para interpretar fenômenos naturais e artificiais de interesse[59]. Ao mesmo tempo, o fato de que a terminologia empregada se manteve razoavelmente preservada a despeito da diversidade das aplicações certamente contribuiu para acelerar a prosperidade da área[7].

No que tange à modelagem computacional de proteínas, o emprego de modelos de rede tem se mostrado fácil de justificar. Do ponto de vista teórico, estruturas de proteínas são naturalmente discretizáveis na escala de resíduos de aminoácidos; não existem mecanismos como inserções, deleções ou trocas de identidade que possam afetar átomos individualmente, e toda a diversidade de estruturas e funções proteicas se apoia nas combinações dos mesmos aminoácidos padrão em números e ordens distintas. Do ponto de vista dos resultados, sucessos obtidos na previsão de propriedades físico-químicas de proteínas a partir de modelos de rede[60–68] sugerem que para muitas aplicações o efeito da topologia sobrepuja, em primeira aproximação, as diferenças entre resíduos, e pouca informação é perdida ao representá-los como vértices idênticos.

Mais surpreendente do que a ampla aplicabilidade das ferramentas provenientes do estudo de redes complexas é a observação de que frequentemente redes construídas para modelar sistemas de origens muito diversas apresentam estruturas e propriedades muito similares[7]. A seguir, formalizaremos os conceitos básicos que permitem *quantificar* essa observação e outras dessa natureza. Cabe ressaltar, contudo, uma distinção semântica. Em [57], Newman refere-se a *redes* e *grafos* como sinônimos, apontando que Teoria dos Grafos é um dos fundamentos da matemática discreta, tendo nascido com a solução do *problema das pontes de Königsberg* por Euler em 1735. Do mesmo modo, Newman não parece oferecer nenhuma característica específica que justifique o adjetivo "complexas", e não diferencia redes complexas de redes em geral. Por outro lado, Costa *et al.* em [69] afirmam que redes complexas são objetos de estudo para a combinação de Teoria dos Grafos e Mecânica Estatística, e que o campo nasce com resultados importantes em percolação e redes aleatórias obtidos por Flory e por Erdős & Rényi em meados do século XX. Assim, para Costa *et al.*, a distinção entre grafos e redes complexas reflete as peculiaridades dos sistemas reais modelados pelas últimas, com padrões estruturais e distribuições características de vértices e conexões que inexistem em modelos mais simples. Aqui, adotamos a segunda acepção, e embora ocasionalmente tratemos redes (complexas) e grafos como sinônimos, ressaltamos que a topologia de contatos em proteínas exibe propriedades típicas de "complexidade", sendo um exemplo a existência de posições privilegiadas ocupadas por resíduos críticos para a manutenção da atividade, uma asserção à qual voltaremos durante toda a seção.

Com isso, oferecemos definições precisas para os conceitos e grandezas com os quais trabalharemos neste capítulo. Seguimos principalmente [7], de onde reproduzimos *verbatim* as definições numeradas, e utilizamos também material adaptado de [3], do autor.

Definição 3.1. *Um grafo G consiste num conjunto V de vértices e um conjunto E de arestas, para o qual denotamos $G = (V, E)$. Diz-se que cada aresta $e \in E$* une *dois vértices. Se e une u,$v \in V$, denota-se $e = \langle u,v \rangle$, e os vértices u e v são ditos adjacentes.*

Matematicamente, grafos são conjuntos de pontos (*nós* ou *vértices*) unidos por

ligações (*arestas*). Embora contidos na definição 3.1, aqui não trataremos de casos em que as extremidades u e v de uma aresta correspondem ao mesmo vértice, isto é, arestas que formam *loops*. Do mesmo modo, não admitiremos a existência de *múltiplas arestas* unindo um mesmo par de vértices. A combinação das duas condições restringe os objetos de estudo ao conjunto dos grafos *simples*. Ressaltamos também que consideramos $\langle u,v\rangle$ equivalente a $\langle v,u\rangle$; arestas dotadas de direcionalidade são uma sofisticação amplamente estudada mas desnecessária para os propósitos deste trabalho.

A iconografia essencialmente universal para grafos consiste em um conjunto de pequenos círculos ou formas geométricas para os vértices unidos por linhas retas para as arestas. A mesma foi introduzida na década de 1930 por Moreno[70], cientista social e psiquiatra e um dos fundadores do campo de redes sociais, para representar a teia de relacionamentos sociais num grupo de crianças pré-escolares, um exemplo do caráter multidisciplinar das aplicações da ciência de redes. A simplicidade da representação traduz a premissa básica que orienta a modelagem de sistemas reais como redes: os nós (bem como as ligações) são representados todos iguais pois são idênticos ou pois suas diferenças são desprezíveis comparadas ao efeito da topologia. De fato, características topológicas importantes podem ficar evidentes pela mera inspeção visual de uma representação gráfica adequadamente construída. Um exemplo é a *planaridade*. A inscrição concreta de um grafo sobre uma superfície dada, atribuindo valores arbitrários porém determinados para as coordenadas da posição de cada vértice, é denominada *imersão*. Grafos planares são aqueles que podem ser imersos no plano sem que exista cruzamento entre suas arestas, isto é, sem que elas se encontrem em pontos diferentes dos vértices[7] (figura 3.2). Para um grafo que represente, por exemplo, uma infraestrutura de transporte urbano, a ausência de planaridade pode implicar na necessidade de soluções verticais como túneis e viadutos, sendo, desta forma, uma exemplo de característica indesejável que pode ser detectada através da análise de uma representação gráfica.

Um segundo exemplo, diretamente relevante no contexto das investigações realizadas, é a identificação de nós que ocupam posições privilegiadas em função da distribuição das ligações – a riqueza de conexões ou o controle das ligações entre nós que estariam de outro modo isolados pode restabelecer hierarquias de importância entre agentes cujas diferenças individuais foram desprezadas. A figura 3.3 ilustra duas medidas intuitivas de importância relativa entre vértices, o número de vizinhos e a participação em muitos caminhos.

Naturalmente, estratégias de análise qualitativas tais como ilustradas nas figuras 3.2 e 3.3 se tornam inadequadas rapidamente à medida que crescem os números de vértices e arestas dos grafos estudados. Assim, é necessário estabelecer interpretações precisas dos atributos de interesse, tais como "posições privilegiadas" ou "comunidades", para as quais possamos definir expressões quantitativas que sejam diretamente calculáveis independente de representações gráficas. Aqui, tendo em consideração o objetivo inicial

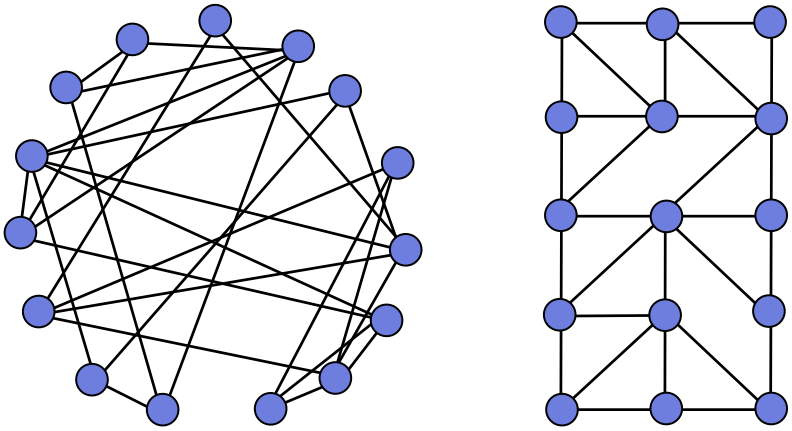


Figura 3.2. Ilustração de duas imersões para o mesmo grafo. À esquerda, uma imersão circular, uma estratégia comumente empregada por garantir que nenhum conjunto de três vértices seja colinear, o que poderia ocasionar ambiguidades na representação de arestas sobrepostas[7], ao mesmo tempo em que é de fácil construção. À direita, uma imersão mais elaborada para o mesmo grafo revela sua planaridade, uma vez que não há cruzamento entre arestas. A mesma também ressalta uma característica intuitiva de "localidade" da rede, na medida em que os vizinhos de cada nó tendem a ser vizinhos entre si, todos os nós participam de números similares de ligações, e inexistem ligações de longa distância. Inspirado em um grafo de [8].

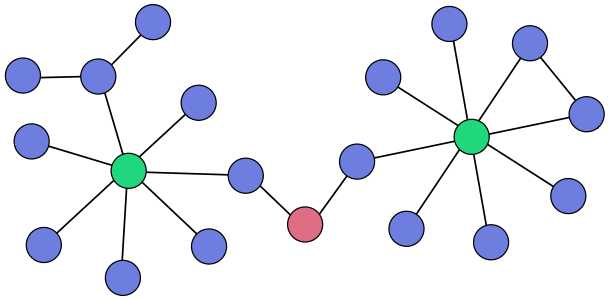


Figura 3.3. Ilustração de grafo com vértices que ocupam posições privilegiadas em termos da distribuição de ligações. Em verde, dois vértices com número de vizinhos muito acima da média para este grafo, atuando como pontos focais locais. Em vermelho, um vértice que, ainda que participe de poucas ligações, consiste na única conexão entre as duas "comunidades" visivelmente delineadas, e sua remoção resultaria na fratura do grafo em componentes separados.

de estudar a relação entre topologia e propagação de calor, enfatizamos atributos que remetam ao potencial de cada nó de impor sua influência sobre o resto da rede. De fato, a importância de um vértice do ponto de vista estrutural pode ser analisada quantitativamente a partir de premissas distintas, originando um conjunto de expressões relacionadas classificadas como medidas de *centralidade*. Introduzidas inicialmente na década de 1950 também no contexto da análise de redes sociais[71, 72], a mais simples dentre elas é

relacionada ao já mencionado número de vizinhos:

Definição 3.2. *Para um grafo G contendo N nós e um vértice $v \in V(G)$, o conjunto $\mathcal{N}(v)$ de vizinhos de v é o conjunto de vértices adjacentes a v, dado por:*

$$\mathcal{N}(v) = \{u \in V(G) \mid u \neq v, \exists e \in E(G) : e = \langle u, v \rangle\}$$

O grau *de v, denotado por $\delta(v)$, é o seu número de vizinhos, $|\mathcal{N}(v)|$.*

A centralidade de grau *de v é dada por:*

$$C_g(v) = \frac{\delta(v)}{N-1} \tag{3.1}$$

Neste ponto, é interessante introduzir a notação da matriz de adjacência, uma representação frequentemente empregada por favorecer alguns tipos de operações em algoritmos que determinam propriedades estruturais de grafos:

Definição 3.3. *Seja G um grafo com N vértices, tal que $V(G) = \{v_1, \ldots, v_N\}$. A* matriz de adjacência $\boldsymbol{A}$ *de G é uma matriz $N \times N$ tal que:*

$$\boldsymbol{A}_{ij} = |\{e \in E(G) \mid e = \langle v_i, v_j \rangle\}|$$

A matriz assim construída descreve completamente o grafo G.

Usando a notação matricial, o grau de um nó é simplesmente a soma da linha ou coluna correspondente na matriz de adjacência, $\delta(v) = \sum_i \mathbf{A}_{vi}$, e, portanto, trivialmente calculado.

Aferir a importância de um vértice pela quantidade de ligações das quais ele participa é uma proposta absolutamente intuitiva, mas padece da deficiência de ser uma medida estritamente local. A figura 3.3 oferece um contraponto com o exemplo de um vértice de centralidade de grau $C_g(v) = 2/19 \sim 0{,}1$, portanto baixa, mas que é todavia nitidamente relevante por pertencer ao único *caminho* que conecta uma fração majoritária dos pares de vértices.

Definição 3.4. *Seja G um grafo. Um* caminho$-(v_0, v_k)$ simples *em G é uma sequência alternante $[v_0, e_1, v_1, e_2, \ldots, v_{k-1}, e_k, v_k]$ de vértices e arestas em G tal que $e_i = \langle v_{i-1}, v_i \rangle$, em que todos os vértices e todas as arestas são distintos. O* comprimento *do caminho$-(v_0, v_k)$ é o número de arestas contidas na sequência alternante.*

Sejam u e v dois vértices em G. Se existe pelo menos um caminho$-(u, v)$, então a distância $d(u,v)$ *entre u e v é o comprimento do menor caminho entre eles.*

A medida que avalia quantitativamente a influência conferida por pertencer a muitos caminhos é a medida da centralidade de intermediação, que mencionamos pelo valor conceitual mas que por brevidade não definiremos formalmente[72]. Ainda assim, o

exemplo revela como uma medida local de importância como o grau pode ser insensível ao efeito global da topologia.

Naturalmente, é possível exercer influência sem ter alto grau ou participar dos caminhos mais curtos dentro da rede, bastando, por exemplo, estar *próximo* a um vértice que figure nestes caminhos; ao invés de buscar muitas conexões diretas ou uma posição privilegiada de intermediação, pode ser suficiente apenas unir-se a um único agente que já ocupe uma posição privilegiada, mas nenhuma das duas medidas citadas reconhece esse mecanismo. Um exemplo de medida que captura esse aspecto da importância do ponto de vista estrutural – a importância dos vértices adjacentes a si – é a centralidade de autovetor, que define a centralidade de cada vértice como a média das centralidades de cada um de seus vizinhos, um sistema de equações autoconsistente que tem a forma de equações de autovetor[73].

Uma maneira alternativa de levar em conta ao mesmo tempo aspectos locais e globais da topologia para avaliar a importância de cada vértice, talvez superior pela simplicidade conceitual, é calcular diretamente a distância média entre o vértice de interesse e cada um dos outros. De fato, é razoável esperar que vértices em posições privilegiadas possam atingir rapidamente todos os outros vértices da rede, seja através seu próprio número de ligações, da participação em caminhos mais curtos entre outros vértices, ou pela proximidade com vértices que exibam tais propriedades, e a minimização da distância média é uma tradução bastante intuitiva do conceito *geométrico* de centro. Definimos, então, a centralidade de proximidade de um vértice como sendo inversamente proporcional à média das distâncias entre ele e todos os outros:

Definição 3.5. *A* centralidade de proximidade *de um vértice v em um grafo G contendo N vértices é dada por:*

$$C_p(v) = (N-1)\left[\sum_{i=1}^{N} d(v,i)\right]^{-1} \tag{3.2}$$

A figura 3.4 ilustra uma comparação visual entre as medidas discutidas.

A literatura revela que medidas de centralidade são frequentemente empregadas com bons resultados para analisar modelos computacionais de proteínas[74–77]. Por brevidade, destacamos apenas [77], em que Amitai *et al.* demonstram que valores altos para a centralidade de proximidade de resíduos de aminoácidos apontam posições conservadas, e podem ser utilizados para identificar sítios funcionais.

Em face dos argumentos apresentados nas seções 2.3, 3.1, e aqui, postulamos que, em redes construídas para representar as interações entre resíduos em estruturas de proteínas, resíduos com alto valor de centralidade devem coincidir com aqueles bons difusores de calor. De fato, sabemos que medidas de centralidade cobrem uma variedade de aspectos diferentes da importância estrutural, e que resultados anteriores sugerem a

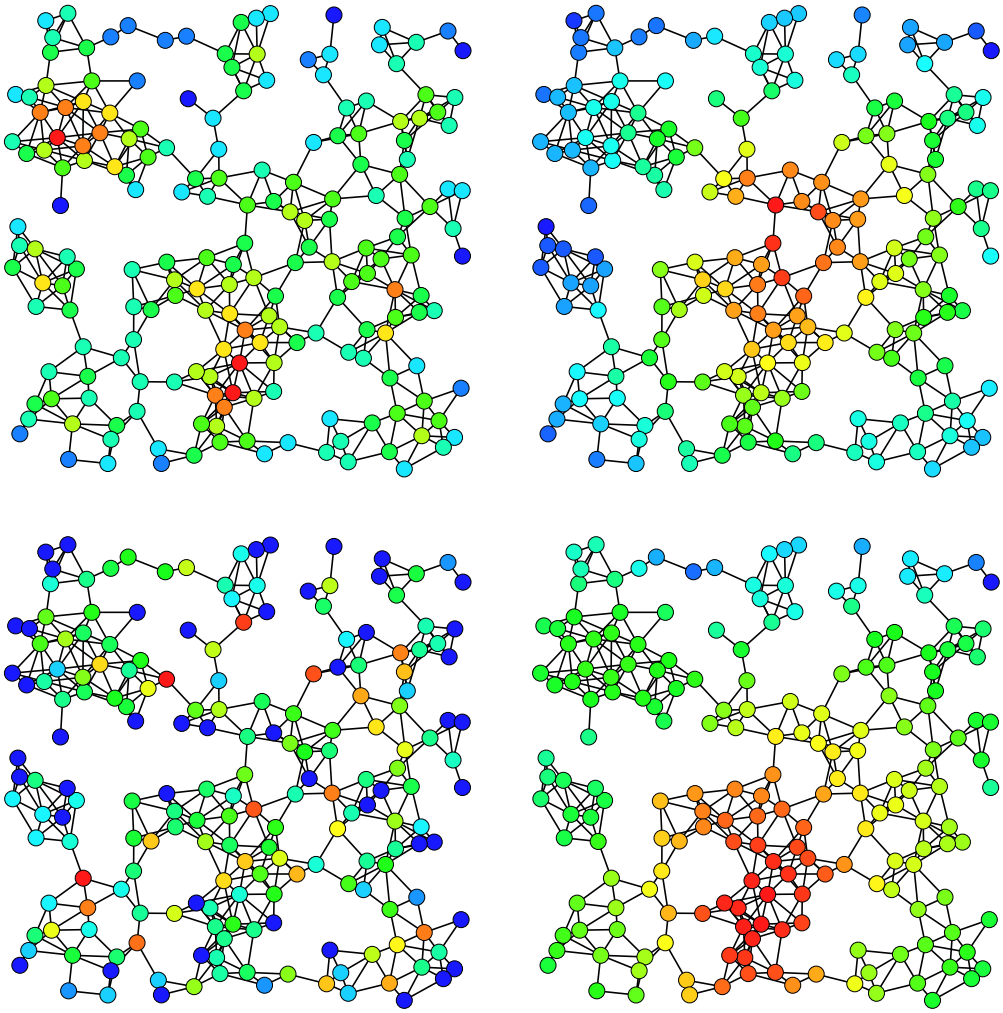


Figura 3.4. Representações do mesmo grafo com vértices colorizados por valores de centralidade calculados segundo medidas distintas. A partir do canto superior esquerdo, em sentido horário: centralidade de grau, centralidade de proximidade, centralidade de autovetor, e centralidade de intermediação. O caráter local das ligações promove a similaridade entre a centralidade de proximidade e a distância euclidiana do centro geométrico da representação. Figura produzida com *software* adaptado de [9].

existência de resíduos com capacidade privilegiada de difundir calor, num processo que é evidentemente mediado pelas interações covalentes e não-covalentes. Assim, algum grau de capacidade preditiva é esperado, e procedemos com essa análise também para saber *qual* aspecto da centralidade melhor traduz a aptidão para transmitir calor. Os resultados dessa análise, obtidos durante o projeto de mestrado do autor, são apresentados a seguir.

Inicialmente, realizamos experimentos com o objetivo de determinar o protocolo adequado para construir redes a partir de estruturas de proteínas. Partimos dos protocolos descritos na literatura, para os quais existe razoável diversidade, e os confrontamos com medidas da distribuição radial de distâncias interatômicas, que calculamos sobre uma grande base de dados estrutural com representantes de todas as classes de eno-

	Coeficiente de correlação de Pearson (r)						
Medida de Centralidade	1F5J	1M4W	1XNB	2VUJ	2VUL	1YS1	2PRG
Centr. de proximidade (C_p)	0,754	0,741	0,723	0,771	0,754	0,746	0,634
Centr. de intermediação (C_i)	0,644	0,603	0,606	0,615	0,625	0,588	0,506
Centr.de grau (C_g)	0,723	0,736	0,717	0,723	0,703	0,746	0,546
Centr. de subgrafo (C_s)	0,708	0,686	0,698	0,724	0,712	0,578	0,542
Centr. de autovetor (C_a)	0,757	0,734	0,724	0,766	0,738	0,689	0,589
Coef. de participação (P)	0,374	0,432	0,349	0,527	0,511	0,527	0,513

Tabela 3.1. Coeficientes de correlação de Pearson entre temperatura final num experimento de ATD por resíduo aquecido e medidas de centralidade por resíduo, para um conjunto de sete proteínas. Os experimentos realizados incluem algumas medidas não descritas no texto, omitidas por serem conceitualmente mais complicadas e terem performance preditiva em geral menor.

velamento conhecidas. As medidas observadas permitiram o descarte de protocolos que empregam definições de "contato" entre resíduos incompatíveis com as escalas de organização interatômica observadas no interior de proteínas. Não apresentaremos aqui estes resultados por se desviarem excessivamente do objetivo da seção, mas mencionamos a análise realizada para assegurar que a metodologia escolhida não é arbitária. O protocolo determinado é apresentado a seguir, e os resultados completos estão disponíveis em [3] ou [43].

Definição 3.6. *Seja uma proteína de N resíduos cuja estrutura é conhecida. A representação desta proteína em forma de rede é obtida construindo sua matriz de adjacência $\boldsymbol{A}$, tal que $\boldsymbol{A}$ é uma matriz $N \times N$ e $\boldsymbol{A}_{ij}$ é nulo a não ser que pelo menos um átomo do resíduo i esteja a no máximo $6\mathring{A}$ de distância de algum átomo do resíduo j, caso em que $\boldsymbol{A}_{ij} = 1$. Diz-se então que os resíduos i e j são* vizinhos*, estão* ligados *ou em* contato*. A matriz assim construída também é denominada* mapa de contatos*.*

Dispondo do protocolo da definição 3.6, construímos redes para representar as estruturas de um conjunto de sete proteínas (códigos PDB: 1F5J, 1M4W, 1XNB, 2VUJ, 2VUL, 1YS1, 2PRG) para as quais dispunhamos de resultados de experimentos de ATD, cedidos por Heloisa Muniz[78] e também pelo orientador do projeto. Para cada uma destas estruturas, calculamos uma série de medidas de centralidade para cada resíduo, e comparamos os resultados com os valores de temperatura final atingida em função do resíduo aquecido em experimentos de ATD. Um exemplo de comparação é apresentado na figura 3.5.

Para cada medida de centralidade e cada proteína, calculamos o coeficiente de correlação de Pearson entre as curvas, e os resultados são apresentados na tabela 3.1.

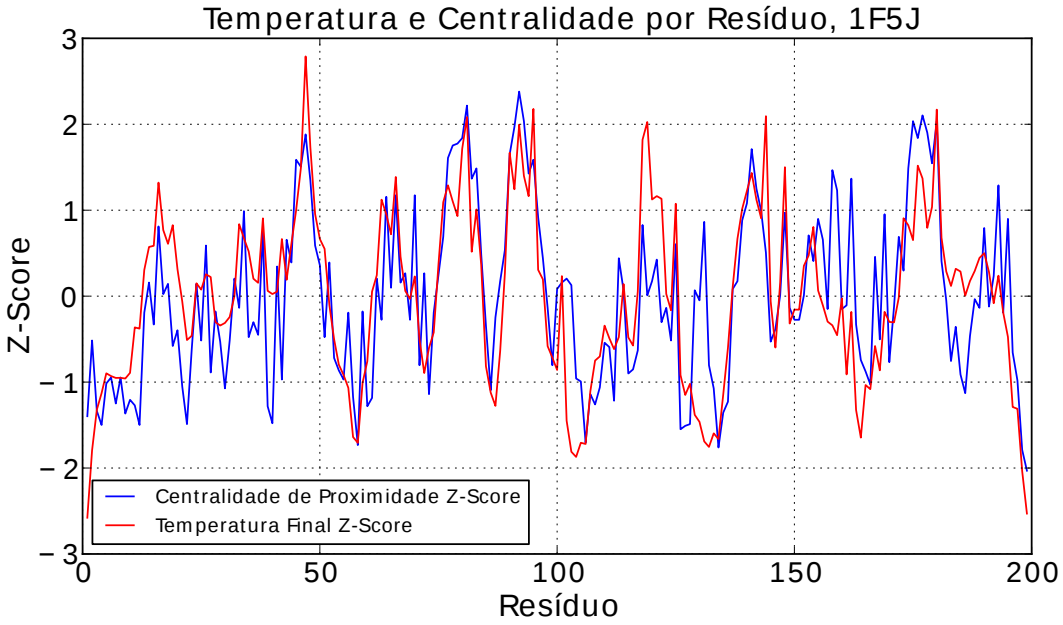


Figura 3.5. Sobreposição entre duas medidas reportadas como função do resíduo ao longo da sequência, para a proteína 1F5J. Em vermelho, a temperatura final atingida após um tempo fixo de aquecimento num experimento de ATD, em função do resíduo aquecido. Em azul, o valor da centralidade de proximidade por resíduo. Ambos os valores são expressados em forma de Z-Score, isto é, em termos do número de desvios padrão em relação à média. A sobreposição entre as duas curvas é notável, tal que o coeficiente de correlação de Pearson é maior que 0,75. Reproduzida de [3], do autor.

Os resultados apresentados encerram o conjunto de experimentos apresentados na dissertação de mestrado do autor, reportados nesta seção para contextualizar o trabalho subsequente.

A inspeção da tabela revela que, conforme esperado, medidas de centralidade são, em maior ou menor grau, capazes de predizer a capacidade de um resíduo de dissipar calor eficientemente sobre o resto da estrutura. Entre as medidas estudadas, destacamos a centralidade de proximidade, que associa a interpretabilidade com os valores mais altos do coeficiente de correlação em quase todos os casos. De fato, por depender apenas das distâncias entre resíduos medidas sobre a rede, um parâmetro eminentemente físico, supusemos que a performance preditiva da centralidade de proximidade se explicaria por esta ser (proporcional a) a própria solução analítica de um modelo físico que reproduza a difusão de calor.

No período coberto por este texto, testamos essa conjectura construindo e resolvendo um modelo analítico para a difusão de calor sobre redes, e estudando o formato funcional da solução para entender a performance de algumas das medidas de centralidade apresentadas. Os resultados são reportados na próxima seção.

3.3 Modelagem analítica da difusão térmica

Durante o período referente a este projeto, retomamos o trabalho realizado com o objetivo de obter uma expressão analítica para a temperatura final da proteína em função do tempo, visando a demonstrar que a correlação observada entre a temperatura final e a centralidade de proximidade é fruto de uma proporcionalidade direta. Escrevemos uma equação de difusão de calor discretizada sobre os nós de uma rede, e obtivemos uma solução analítica para o caso em que um dos nós é mantido acoplado a um banho térmico de temperatura fixa. Apresentamos a demonstração a seguir, corroborando resultados obtidos por Szalay & Csermely em [79] e seguindo o texto de [43], do autor.

Seja $\mathbf{A}$ a matriz de adjacência de resíduos que representa uma dada estrutura. Para o resíduo de índice i, a variação total de temperatura no tempo t depende da diferença de temperatura entre i e cada um de seus vizinhos:

$$\frac{dT_i(t)}{dt} = k \sum_j \mathbf{A}_{ij}(T_j(t) - T_i(t)) \tag{3.3}$$

Onde o papel da constante k é análogo à difusividade térmica, que consideramos neste momento idêntica entre todos os pares de resíduos. Ao escrever a equação 3.3, assumimos que regiões de dimensões similares a um resíduo de aminoácido compreendem um número de átomos suficiente para que se possa falar em equilíbrio térmico, mas não suficiente para que o equilíbrio não se estabeleça rapidamente, de tal forma que o conceito de temperatura "local" do resíduo i no tempo t seja bem definido. Esta premissa não é inédita, e encontra embasamento tanto em medidas da escala de tempo da termalização em proteínas[45, 46] como em trabalhos de modelagem anteriores bem-sucedidos. Separando os termos da equação 3.3, vem:

$$\frac{dT_i(t)}{dt} = k \sum_j \mathbf{A}_{ij} T_j(t) - kT_i(t) \sum_j \mathbf{A}_{ij}$$

Onde, conforme mencionado na seção 3.2 (definição 3.3), a soma $\sum_j \mathbf{A}_{ij}$ da linha i da matriz de adjacência é o grau do vértice i, $\delta(i)$. Obtemos:

$$\frac{dT_i(t)}{dt} = k \sum_j \mathbf{A}_{ij} T_j(t) - kT_i(t)\delta(i)$$

Reunindo as temperaturas $T_i(t)$ na forma de vetor coluna, vem:

$$\frac{d\mathbf{T}(t)}{dt} = k\mathbf{A}\mathbf{T}(t) - k\mathbf{D}\mathbf{T}(t) = -k(\mathbf{D} - \mathbf{A})\mathbf{T}(t) = -k\mathbf{L}\mathbf{T}(t) \tag{3.4}$$

Onde $\mathbf{D} = \mathrm{diag}\{\delta(1), \delta(2), \ldots, \delta(N)\}$ é a matriz de graus para este grafo, uma matriz diagonal cujos elementos correspondem aos graus de cada vértice dispostos sobre a diagonal principal, e a matriz $\mathbf{L} = \mathbf{D} - \mathbf{A}$ é denominada matriz *Laplaciana* do grafo.

A equação diferencial 3.4 tem solução em termos da decomposição de $\mathbf{L}$ em autovetores, que também pode ser dada em formato de matriz exponencial:

$$\mathbf{T}(t) = e^{-\mathbf{L}kt}\mathbf{T}(0) = \sum_j \{\mathbf{v}_j^T \cdot \mathbf{T}(0)\} e^{-\lambda_j kt}\mathbf{v}_j \tag{3.5}$$

Na equação 3.5, $\mathbf{v}_j$ são os autovetores de $\mathbf{L}$ e λ_j são os autovalores correspondentes. Em particular, notamos que o autovetor $\mathbf{v}_0 = \mathbf{1}$, associado ao menor autovalor $\lambda_0 = 0$, sempre é solução, pois todas as linhas e colunas de $\mathbf{L}$ tem soma zero. Neste ponto, podemos simplificar a solução incluindo características particulares do sistema de interesse. Considerando que a configuração inicial de uma simulação de ATD corresponde à estrutura resfriada a uma temperatura muito baixa, e que durante a simulação apenas um resíduo é acoplado ao reservatório térmico à temperatura ambiente, podemos simplificar a solução impondo que o vetor de temperaturas iniciais contenha apenas um componente não-nulo. Fazendo $\mathbf{T}(0) = [\ldots, 0, \theta, 0, \ldots]^T$, onde θ é a temperatura inicial do resíduo h aquecido, vem:

$$\mathbf{T}(t) = \theta \sum_j [\mathbf{v}_j]_h e^{-\lambda_j kt}\mathbf{v}_j \tag{3.6}$$

Em que $[\mathbf{v}_j]_h$ é o h-ésimo componente do j-ésimo autovetor. A temperatura de cada resíduo será:

$$T_i(t) = \theta \sum_j [\mathbf{v}_j]_h e^{-\lambda_j kt}[\mathbf{v}_j]_i$$

Sob a equação 3.5, para tempos longos o vetor das temperaturas $\mathbf{T}(t)$ sempre reduz-se a um vetor constante, proporcional ao autovetor $\mathbf{v}_0 = \mathbf{1}$, que corresponde à situação de equilíbrio térmico entre todos os resíduos. Por esta razão, a evolução sob a equação 3.6 se assemelha à relaxação de um pulso concentrado. Assim, para possibilitar a comparação com dados de ATD, pode-se tomar como parâmetro o tempo até o equilíbrio como função do resíduo aquecido, ou a taxa inicial com que a energia deixa o resíduo aquecido. Aquecendo o resíduo h, podemos aproximar, para $kt \ll 1$:

$$T_h(t) = \theta \sum_j e^{-\lambda_j kt}[\mathbf{v}_j]_h^2 = \theta \sum_j (1 - \lambda_j kt + \mathcal{O}((kt)^2))[\mathbf{v}_j]_h^2$$

Desprezando termos da ordem de $(kt)^2$ e mais altos:

$$T_h(t) \approx \theta(\sum_j [\mathbf{v}_j]_h^2 - kt \sum_j \lambda_j [\mathbf{v}_j]_h^2)$$

Resultando:

$$T_h(t) \approx \theta(1 - \mathbf{L}_{hh}kt) = \theta(1 - \delta(h)kt) \tag{3.7}$$

Pela equação 3.7 a taxa de dissipação inicial é, em primeira ordem, função do grau do resíduo aquecido – resíduos com mais vizinhos levam a dinâmica mais rapidamente para o equilíbrio. Esta observação remete ao resultado reportado por Moreno & Pacheco[80], que demonstram que quando uma rede livre de escala de osciladores acoplados sofre uma perturbação local, o tempo médio para o retorno à sincronização de fases é uma função do grau k do nó perturbado, com expoente $\langle\tau\rangle \sim k^{-0,96}$ muito próximo a -1, embora o argumento apresentado como justificativa seja dirigido a topologias de rede em que ciclos fechados são infrequentes.

De fato, observamos uma correlação negativa entre o tempo até o equilíbrio em função do resíduo aquecido, dado pela equação 3.6, e a temperatura final em função do resíduo aquecido, resultados que não apresentaremos aqui. Contudo, a evolução sob a equação 3.5 preserva a temperatura média, $(1/N)\sum_i \mathbf{T}_i(t) = (1/N)\sum_i \mathbf{T}_i(0), \forall t$, também consequência da soma zero das linhas e colunas da matriz Laplaciana. Essa propriedade sugere que a equação 3.5 é inadequada para modelar um experimento de ATD, no qual o resíduo de interesse é mantido acoplado ao banho térmico durante toda a duração da simulação e a temperatura média cresce monotonicamente.

Podemos incluir um termo para o reservatório térmico na equação 3.3 para corrigir o modelo:

$$\frac{dT_i(t)}{dt} = k_b\mathbf{B}_{ii}(\theta - T_i) + k\sum_j \mathbf{A}_{ij}(T_j(t) - T_i(t)) \tag{3.8}$$

Onde θ denota a temperatura do banho térmico, que coincide com a temperatura inicial do resíduo aquecido. A matriz $\mathbf{B}$ é uma matriz diagonal construída tal que $\mathbf{B}_{ii}$ é 1 se o resíduo i está acoplado ao reservatório térmico, e zero do contrário. Para resolver a equação 3.8, introduzimos a substituição $T_i^{\text{rel}} = T_i - \theta$, fazendo:

$$\frac{dT_i^{\text{rel}}(t)}{dt} = -k_b\mathbf{B}_{ii}T_i^{\text{rel}} + k\sum_j \mathbf{A}_{ij}(T_j^{\text{rel}}(t) - T_i^{\text{rel}}(t))$$

Em forma de vetor coluna:

$$\frac{d\mathbf{T}^{\text{rel}}(t)}{dt} = -k_b\mathbf{B}\mathbf{T}^{\text{rel}}(t) - k\mathbf{L}\mathbf{T}^{\text{rel}}(t) = (-k\mathbf{L} - k_b\mathbf{B})\mathbf{T}^{\text{rel}}(t)$$

A equação obtida é análoga à equação 3.4, e a resolvemos usando a mesma técnica:

$$\mathbf{T}^{\text{rel}}(t) = e^{(-\mathbf{L}kt - \mathbf{B}k_bt)}\mathbf{T}^{\text{rel}}(0) = e^{-\mathbf{M}kt}\mathbf{T}^{\text{rel}}(0) \tag{3.9}$$

Que resulta em:

$$\mathbf{T}(t) = e^{-\mathbf{M}kt}\mathbf{T}(0) + (\mathbf{I} - e^{-\mathbf{M}kt})\boldsymbol{\theta} \tag{3.10}$$

	Proteína						
Parâmetros do modelo	1F5J	1M4W	1XNB	2VUJ	2VUL	1YS1	2PRG
τ	15,0	18,0	9,5	14,5	11,0	17,0	32,0
k_r	10^3	10^3	10^3	10^3	10^3	10^3	10^3
Coef. de Pearson r	0,772	0,774	0,751	0,775	0,749	0,793	0,627

Tabela 3.2. Coeficientes de correlação de Pearson entre temperatura final num experimento de ATD por resíduo aquecido e a mesma medida calculada pela equação 3.10, para o mesmo conjunto de proteínas da seção anterior. Apresentamos os parâmetros que maximizam as correlações observadas em cada caso. Os coeficientes de correlação deixam de variar se o valor de k_r cresce para além do indicado, sugerindo que 10^3 seja grande o suficiente para caracterizar um equilíbrio com o banho térmico essencialmente instantâneo se comparado ao acoplamento entre resíduos. No caso de τ, os valores apresentam variação relativa maior, mas são também todos da mesma ordem de grandeza.

O símbolo $\boldsymbol{\theta}$ denota o vetor $[\ldots, \theta, \theta, \theta, \ldots]^T = \theta\mathbf{1}$. Evoluindo sob a equação 3.10, a temperatura média aumenta monotonicamente e $\mathbf{T}(t)$ se reduz a $\boldsymbol{\theta}$ para tempos longos, reproduzindo o comportamento de simulações de ATD. A matriz $\mathbf{M}$ é dada por $\mathbf{M}kt = \mathbf{L}kt + \mathbf{B}k_bt = (\mathbf{L} + k_r\mathbf{B})\tau$, e depende de dois parâmetros: $\tau = kt$ é o tempo característico do acoplamento entre resíduos, e $k_r = k_b/k$ é a intensidade relativa do acoplamento com o banho térmico comparado ao acoplamento entre resíduos. A matriz $\mathbf{M}$ é uma perturbação da matriz Laplaciana que inclui a influência do reservatório térmico.

Para avaliar a solução obtida, comparamos a temperatura final calculada segundo a equação 3.10 com os dados de ATD utilizados na seção anterior, e os resultados são apresentados na tabela 3.2 e ilustrados nas figuras 3.6 e 3.7.

As correlações apresentadas são ligeiramente melhores que aquelas obtidas para as medidas de centralidade na seção anterior, e o sucesso da modelagem é corroborado pela ótima sobreposição entre as curvas teóricas e as medidas de ATD. Tendo em vista a simplicidade do modelo, a capacidade da solução encontrada de descrever a dissipação de energia é notável; antecipamos que melhoras na performance só possam ser obtidas com a introdução de sofisticações, como por exemplo a variação das constantes de difusividade térmica em função da natureza da interação entre resíduos.

Por outro lado, a expectativa inicial que motivou esta análise era de que a solução analítica do modelo revelaria uma proporcionalidade direta com a medida de centralidade de proximidade (equação 3.2), e essa não se concretizou. Investigamos brevemente se o formato funcional da solução encontrada correspondia a alguma medida já descrita na literatura. Como na seção anterior, buscamos entender o comportamento da

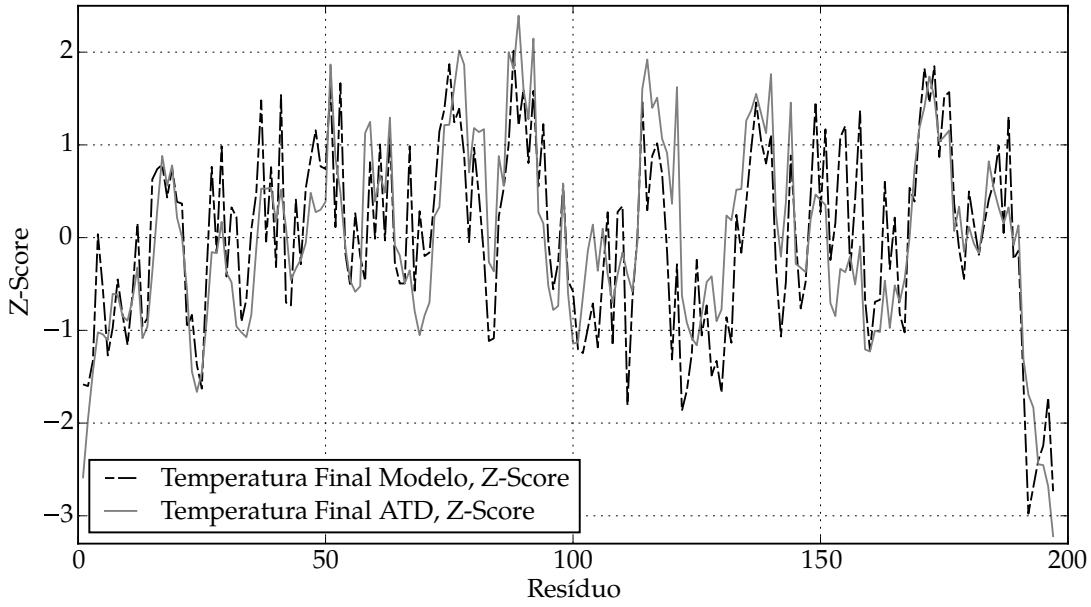


Figura 3.6. Sobreposição entre duas medidas dadas em função da posição ao longo da sequência, para a proteína 1M4W. Em cinza, com linha contínua, a temperatura final atingida após um tempo fixo de aquecimento num experimento de ATD, como função do resíduo aquecido. Em preto, com linha tracejada, a mesma medida calculada pela equação 3.10. Os valores são reportados na forma de Z-Score, *i. e.*, distância da média expressada em número de desvios padrão.

taxa inicial de dissipação, aproximando até segunda ordem para $\tau \ll 1$:

$$\mathbf{T}(\tau) - \mathbf{T}(0) \approx (-\mathbf{M}\tau + \mathbf{M}^2\frac{\tau^2}{2})\mathbf{T}(0) + (\mathbf{M}\tau - \mathbf{M}^2\frac{\tau^2}{2})\boldsymbol{\theta}$$

Tomando a média, vem:

$$\frac{1}{N}\sum \Delta\mathbf{T}(\tau) \approx \frac{\tau}{N}[\sum \mathbf{M}\boldsymbol{\theta} - \sum \mathbf{M}\mathbf{T}(0)] + \frac{\tau^2}{2N}[\sum \mathbf{M}^2\mathbf{T}(0) - \sum \mathbf{M}^2\boldsymbol{\theta}]$$

O índice h denota o resíduo aquecido. Separando os termos aditivos para facilitar o cálculo, lembramos que as linhas e colunas de $\mathbf{L}$ tem soma zero, de forma que a ação de $\mathbf{M}$ sobre $\boldsymbol{\theta}$ equivale a soma das linhas de $k_r\mathbf{B}$ e resulta $k_r\theta$. Do mesmo modo, a ação de $\mathbf{M}$ em $\mathbf{T}(0)$ é a h-ésima coluna de $k_r\mathbf{B}$ e também soma $k_r\theta$.

Para calcular os termos quadráticos em $\mathbf{M}$, lembramos que $\mathbf{L}$ e $\mathbf{B}$ não comutam, e aplicamos $\mathbf{M}^2 = (\mathbf{L}^2 + k_r\mathbf{LB} + k_r\mathbf{BL} + k_r^2\mathbf{B}^2)$. Os termos que dependem de $\mathbf{L}^2$ tem soma nula, assim como os termos que dependem de $\mathbf{LB}$. O termo $\sum \mathbf{BLT}(0)$ vale $\delta(h)\theta$, e os termos que dependem de $\mathbf{B}^2$ somam $\sum \mathbf{B}^2\boldsymbol{\theta} = \theta$ e $\sum \mathbf{B}^2\mathbf{T}(0) = \theta$. Introduzimos na

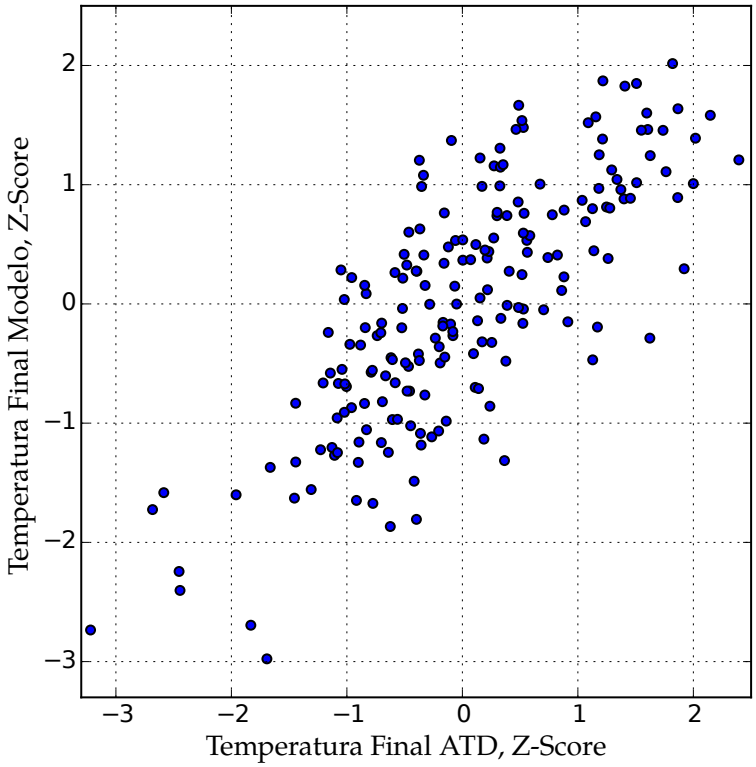


Figura 3.7. Correlação entre temperatura final atingida em função do resíduo aquecido e a mesma medida calculada pela equação 3.10, para a estrutura 1M4W. Os valores são dados na forma de Z-Score, a distância da média expressada em número de desvios padrão.

média, obtemos:

$$\Delta T_{\text{avg}}(\tau) \approx \frac{\tau}{N}[k_r\theta - k_r\theta] + \frac{\tau^2}{2N}[k_r\theta\delta(h) + k_r^2\theta - k_r^2\theta]$$

Finalmente:

$$\Delta T_{\text{avg}}(\tau) \approx \frac{\tau^2}{2N}k_r\theta\delta(h)$$

Resultando que a taxa inicial de crescimento da temperatura média é proporcional, até segunda ordem, à temperatura do banho térmico, à intensidade do acoplamento ao banho, e ao grau do resíduo aquecido, que lembramos ser uma das medidas de centralidade com boa performance preditiva.

Para tempos mais longos, todavia, a aproximação deixa de ser válida, e é necessário inspecionar as equações 3.9 ou 3.10 em busca de uma constante de proporcionalidade. Lamentavelmente, os resultados do ajuste do modelo (tabela 3.2) indicam que a perturbação adicionada à matriz Laplaciana pelo banho térmico para construir a matriz $\mathbf{M} = \mathbf{L} + k_r\mathbf{B}$ é absolutamente não desprezível, de forma que o comportamento da matriz exponencial $e^{-\mathbf{M}kt}$ não é trivial de aproximar a partir do comportamento razoavelmente bem estudado de $e^{-\mathbf{L}}$. Não obstante, cabe mencionar que, sob

a equação 3.10, a evolução da temperatura média tem o formato $(1/N)\sum_i \mathbf{T}_i(t) = \theta - (\theta/N)\left(\sum_i \left[e^{-\mathbf{M}kt}\mathbf{1}\right]_i - \sum_i \left[e^{-\mathbf{M}kt}\right]_{hi}\right)$. O termo de sinal negativo que multiplica (θ/N) é a constante de proporcionalidade direta que cumpre o papel de medida de centralidade do resíduo aquecido h, e consiste na soma de todos os termos da matriz $e^{-\mathbf{M}kt}$ menos a soma da linha (ou coluna) h da mesma matriz. Como a própria matriz $\mathbf{M}$ depende de quem é o resíduo aquecido, a soma dos termos de $e^{-\mathbf{M}kt}$ não é uma invariante da rede (ao contrário de $e^{-\mathbf{L}}$) e não pode ser ignorada. Ainda assim, a soma da linha h da matriz $e^{-\mathbf{M}kt}$ remete à medida de *comunicabilidade total*[73, 81], uma medida de centralidade que tem formato similar porém como função da matriz de adjacência A.

Os resultados apresentados neste capítulo estão publicados na revista *Bioinformatics* em [43].

Capítulo 4

Complexidade topológica e cinética de enovelamento

O trabalho descrito no capítulo 3 encerrou as análises de difusão térmica em proteínas. Entretanto, os conceitos e a metodologia desenvolvidos para a modelagem de estruturas oferecem uma perspectiva de análise cuja aplicabilidade certamente é mais ampla e se estende a problemas de natureza distinta. Em particular, descrições que enfatizam a topologia das interações aparentam ser convenientes para estudar problemas que requeiram a descrição da *complexidade* de estruturas segundo alguma definição.

Aqui, empregamos com esse objetivo uma abstração similar à desenvolvida no trabalho anterior, estudando a relação entre a cinética da reação de enovelamento em proteínas e a complexidade das estruturas nativas correspondentes. Para tanto, apresentamos na seção 4.1 uma discussão concisa da reação de enovelamento em proteínas. Ao longo da seção 4.1 e do resto deste capítulo, seguimos o texto de [82], do autor.

4.1 A reação de enovelamento

Na seção 1.1, mencionamos *en passant* o caráter altamente *reprodutível* do conjunto de estruturas que define o estado nativo de uma proteína, no sentido de que sob condições fisiológicas as configurações adotadas pela cadeia são previsíveis e dependem apenas de sua sequência de aminoácidos. A elegante demonstração experimental deste argumento é creditada a Anfinsen[83], por um conjunto de estudos do enovelamento da ribonuclease pancreática bovina cujo início, na década de 1950, precedeu inclusive a publicação da primeira estrutura tridimensional de uma proteína.

A delicada dependência da manutenção da estrutura nativa com a estabilidade das condições físico-químicas do ambiente intracelular (ou da solução) é uma condição facilmente observada na bancada do laboratório, onde o trato cotidiano com proteínas frequentemente envolve o controle agressivo das condições experimentais para preservar a atividade das soluções. Ao mesmo tempo, a desnaturação nao é necessariamente um processo irreversível. A depender da proteína e da natureza da perturbação que ocasionou a perda da estrutura, o restabelecimento das condições iniciais pode bastar para restaurar a estrutura nativa e, com ela, a atividade.

Anfinsen e colaboradores demostraram contundentemente que, no caso da ribonuclease pancreática bovina, a capacidade de recuperar o enovelamento correto não se devia à persistência das pontes dissulfeto, ligações covalentes entre resíduos distantes que poderiam sobreviver a condições desnaturantes suaves. De fato, aplicando e removendo ureia (um desnaturante) e um agente redutor (para clivar as pontes dissulfeto) em ordens variadas, Anfinsen e colaboradores mostraram[84] que era possível "embaralhar" as pontes dissulfeto, resultando numa mistura de estruturas com atividade drasticamente reduzida por incluir apenas uma pequena fração de moléculas corretamente enoveladas. Removendo os agentes na ordem oposta, era possível recompor primeiro a estrutura terciária correta e só então as pontes dissulfeto, e com isso recuperar a quase totalidade da atividade inicial. Esta observação foi corretamente interpretada como evidência de que toda a informação necessária para levar à estrutura nativa, pelo menos no caso de proteínas globulares pequenas, está contida na estrutura primária, inexistindo a necessidade de enovelamento gradual à medida que a cadeia é sintetizada ou da participação de maquinário bioquímico específico que induza ao enovelamento segundo algum gabarito. A este paradigma, a noção de que o enovelamento é consequência apenas das interações entre os resíduos da cadeia, denominou-se *hipótese termodinâmica*.

Em contrapartida, essa propriedade não implica que a toda sequência de aminoácidos corresponde uma estrutura enovelada bem definida, e de fato a imensa maioria das sequências não leva a estrutura alguma. A capacidade de enovelar previsivelmente é subordinada a considerações dinâmicas e cinéticas bastante restritivas, de forma que as estruturas primárias de proteínas naturais são significativamente diferentes de sequências aleatórias quando analisadas em termos da distribuição das propriedades físico-químicas dos resíduos[85], uma observação que certamente reflete o resultado da atuação da seleção natural ao longo de muitas gerações.

Do ponto de vista termodinâmico, o mecanismo que induz ao enovelamento é consequência da necessidade de compensar a perda de entropia associada à adoção de um conjunto restrito de conformações. Essa diminuição, desfavorável, é amortizada por dois tipos de fenômeno. O primeiro é a formação de interações energeticamente favoráveis que decorrem da aproximação entre resíduos previamente distantes; interações eletrostáticas importantes entre pares de resíduos carregados, por exemplo, mas também grande número de ligações de hidrogênio envolvendo átomos do backbone e também das cadeias laterais dos resíduos polares e carregados. Contudo, a variação da entalpia associada à formação dessas interações não é em geral suficiente para contrapor a diminuição da entropia, em grande medida pelo fato de que muitos destes mesmos doadores e aceptores de ligações de hidrogênio se satisfazem interagindo com moléculas de solvente no estado desenovelado.

O segundo fenômeno, talvez mais importante, se deve à *ausência* de ligações de hidrogênio entre as moléculas de solvente e as cadeias laterais de resíduos hidrofóbicos no estado desenovelado. A perturbação da cadeia de ligações de hidrogênio entre moléculas

de água, devida à impossibilidade de solvatar os grupos apolares, induz a diminuição da área de superfície hidrofóbica exposta ao solvente, mediante a aproximação entre pares de resíduos hidrofóbicos. Este efeito, denominado *efeito hidrofóbico*, pode se originar tanto do reestabelecimento de ligações de hidrogênio água-água, um efeito entálpico, ou da diminuição na entropia rotacional das moléculas de água forçosamente organizadas em torno da superfície hidrofóbica exposta ao solvente[86, 87]. O fenômeno resultante, observado em essencialmente todas as proteínas globulares, é denominado *colapso do núcleo hidrofóbico*, e é ilustrado na figura 4.1.

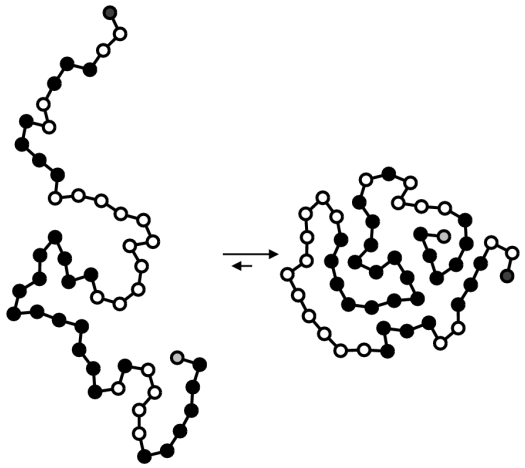


Figura 4.1. Ilustração bidimensional do enovelamento de uma cadeia peptídica impelido pelo colapso do núcleo hidrofóbico. Resíduos hidrofílicos são representados em branco, e hidrofóbicos em preto. A diminuição da superfície hidrofóbica total exposta ao solvente é um mecanismo importante de favorecimento do enovelamento. Obtida de [10].

Finalmente, embora a contribuição do colapso do núcleo hidrofóbico e das interações favoráveis seja importante para equilibrar o custo entrópico do enovelamento, apontamos que o efeito das interações intramoleculares é tal que a reação não é necessariamente mais favorável quanto maior o número de pares de resíduos interagentes. A distribuição das interações no espaço, e, consequentemente, a distribuição dos resíduos carregados, hidrofílicos e hidrofóbicos ao longo da sequência, obedece a uma restrição adicional sutil. Resíduos que formam ligações de hidrogênio ou interações eletrostáticas em geral não são capazes de formá-las apenas com o par pretendido na estrutura correta, e a formação de ligações que não existem na estrutura nativa configura uma "armadilha" que pode retardar significativamente a reação de enovelamento – situação na qual a superfície de energia livre é dita excessivamente "rugosa", e as múltiplas possibilidades de interações são ditas "frustradas"[24]. Além disso, uma sequência que admite a formação de interações compatíveis com múltiplas configurações concorrentes não é robusta em relação a mutações aleatórias, pois, no limite, uma única mutação pontual pode levar uma conformação inativa a se tornar o novo mínimo global de energia livre e, com isso, a

estrutura mais provável[24]. Por estas razões, a cooperatividade e reforço mútuo das interações presentes na estrutura nativa é alvo de pressão seletiva, para que o enovelamento se complete de maneira previsível e dentro de um tempo razoável do ponto de vista de sua função biológica. A esta restrição, um exemplo de influência cruzada entre aspectos cinéticos e dinâmicos do enovelamento, se denomina *princípio da frustração mínima*[24].

O aspecto cinético da reação de enovelamento, por sua vez, tem sido objeto de investigação desde o importante seminário em que Levinthal[88] destacou o número astronômico de estruturas possíveis para uma sequência de tamanho típico, e a consequente impossibilidade de encontrar o enovelamento correto em um tempo razoável sem o emprego de um mecanismo eficiente de amostragem de conformações. No mesmo discurso, Levinthal aventou a hipótese de que o dito mecanismo depende da constituição rápida de interações *locais*, isto é, entre resíduos próximos ao longo da cadeia, que subsequentemente atuam como pontos de nucleação para promover e acelerar o enovelamento[88].

Em [89], Karplus & Weaver argumentam que o mecanismo proposto por Levinthal é precursor direto do modelo de Nucleação-Condensação (NC) para o enovelamento. O modelo NC é caracterizado pela ausência de organização observável antes da formação do estado de transição, seja na forma de estrutura secundária ou de colapso do núcleo hidrofóbico. A formação do estado de transição, que exibe rudimentos tanto de estrutura secundária quanto terciária, constitui então a etapa limitante da taxa de reação[90]. No mesmo trabalho, Karplus & Weaver introduzem o mecanismo de Difusão-Colisão (DC), modelando o enovelamento como um processo hierárquico cuja primeira etapa é a formação rápida de "microdomínios" de caráter local e estabilidade limítrofe, que empreendem movimentos difusivos, colidindo entre si até coalescer. O modelo DC é, em tese, compatível com a existência de uma pluralidade de caminhos de enovelamento concorrentes, que refletem as diferentes ordens em que os microdomínios formados podem coalescer, ao passo que no modelo NC o enovelamento passa pela formação obrigatória de pontos de nucleação bem definidos, sobre os quais a formação simultânea de estruturas secundárias e terciárias é favorecida[89].

Na literatura, é possível encontrar argumentos considerando cada um dos modelos como sendo o mais geral, enquanto o modelo oposto é tratado como um caso particular observado apenas sob regimes específicos[89, 91]. Do ponto de vista do autor, ambos os modelos tal qual definidos são compatíveis com a descrição (algo inespecífica) oferecida por Levinthal. Ademais, já se mostrou possível em algumas proteínas modular o mecanismo de enovelamento através da variação das condições experimentais ou por meio de mutações sítio-dirigidas, preservando a estrutura do estado nativo enquanto o mecanismo de enovelamento explora o espectro NC-DC[92]. Na prática, a estabilidade intrínseca de trechos locais de estrutura secundária na ausência de interações terciárias aparenta ser o principal parâmetro que determina se o mecanismo de enovelamento se parece mais com o modelo NC ou DC[91].

Do ponto de vista experimental, historicamente as investigações da cinética de enovelamento têm privilegiado a medida das taxas de desnaturação (*unfolding rate*, k_u) e reenovelamento (*folding rate*, k_f) em função das condições físico-químicas do ambiente[93, 94], em geral mediante a adição controlado de agentes desnaturantes como ureia ou cloreto de guanidínio. Em muitas proteínas pequenas, ambas as taxas variam linearmente com a concentração de desnaturante; esta observação é tipicamente interpretada como evidência de que a reação progride sem a formação de intermediários estáveis. Esta conjectura, por sua vez, pode ser confirmada através da comparação entre uma medida direta e independente da variação de energia livre de desnaturação em condições de equilíbrio e a estimativa $\Delta G_U = -RT \ln(k_f/k_u)$ calculada a partir das taxas cinéticas[93]. Quando o enovelamento é bem descrito por esta linearidade, o perfil da reação é denominado enovelamento "de dois estados" ou "*two-state*", e a taxa de enovelamento na ausência de desnaturante pode ser obtida pela extrapolação das taxas medidas em concentrações finitas.

Em outras proteínas pequenas, muitas proteínas maiores e mesmo algumas que enovelam segundo o perfil two-state em condições distintas, observa-se que a taxa de reenovelamento atinge um patamar conforme a concentração de desnaturante se aproxima de zero, um desvio da linearidade que dificulta a sua medição por extrapolação. Considera-se que este fenômeno, conhecido por *rollover* (algo como "capotagem") no gráfico de folding rate contra concentração de desnaturante[95], sinaliza a existência de intermediários estáveis de enovelamento[96], um perfil de reação denominado enovelamento "multiestado" ou "*multistate*". Exemplos dos dois perfis são ilustrados na figura 4.2.

De modo geral, condições que favorecem significativamente o estado enovelado podem estabilizar também os intermediários de enovelamento e, com isso, alterar o perfil de enovelamento de two-state para multistate, um efeito reminiscente das "armadilhas" de enovelamento discutidas em relação ao princípio da frustração mínima. Inversamente, condições que promovem a cooperatividade das interações, tais como a inclusão de termos de longa distância ou de muitos corpos em campos de força simulados, podem induzir o efeito oposto[95, 97].

Em vista das observações relatadas, pode-se intuir alguma correlação entre os mecanismos propostos para o enovelamento, de Nucleação-Condensação ou Difusão-Colisão, e os perfis experimentais, two-state ou multistate; averiguar rigorosamente essa relação, contudo, foge ao escopo do trabalho apresentado. Em vez disso, nos concentramos aqui nos esforços de predição teórica das taxas de reação. De fato, os mesmos modelos também parecem sugerir uma distinção funcional importante entre contatos *locais* e contatos *de longa distância*. Os primeiros são muitas vezes formados espontaneamente dentro de intervalos muitas ordens de grandeza menores que o tempo de enovelamento[89], enquanto a formação dos segundos por vezes caracteriza a formação do estado de transição de

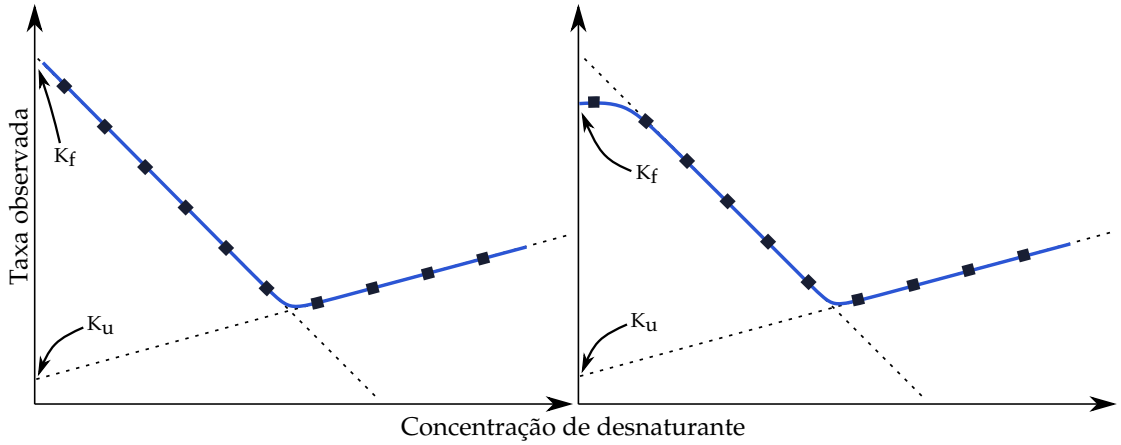


Figura 4.2. Ilustração do formato típico do gráfico do folding rate, em escala logarítmica, *versus* a concentração de desnaturante para experimentos de desnaturação ou reenovelamento induzidos. O gráfico é comumente denominado *"chevron plot"* na literatura, fruto da similaridade entre seu formato em "V" e o formato típico de insígnias militares de mesmo nome. À esquerda, um perfil típico para uma proteína com enovelamento two-state; o folding rate é facilmente extrapolado para uma condição de ausência de desnaturante. À direita, um perfil típico de um enovelamento multistate, com um exemplo de "rollover" na vizinhança da concentração igual a zero dificultando a extrapolação linear da taxa de enovelamento.

enovelamento [98]. Na próxima seção, investigamos uma consequência desta observação, na forma de uma relação entre a prevalência de contatos de longa distância na estrutura enovelada e a taxa de enovelamento experimental.

4.2 Modelos preditivos do folding rate

Os esforços para modelar a reação de enovelamento com o objetivo de reproduzir medidas experimentais resultaram na identificação de descritores estruturais bastante simples que exibem correlações significativas com as taxas observadas. Talvez não surpreendentemente, o mais direto entre eles é o próprio comprimento da cadeia, embora esta correlação específica seja bastante dependente do mecanismo de enovelamento[99]. Mais notavelmente, uma relacão significativa entre a *topologia* da estrutura nativa e o logaritmo da taxa de enovelamento foi estabelecida relativamente cedo, capturada na medida da "ordem de contato" ou *contact order* da estrutura nativa[100–104]. Introduzida como *relative contact order*, a medida é uma estimativa da complexidade da topologia do estado enovelado, calculada em termos da separação média entre resíduos em contato, tomada ao longo da sequência:

$$RCO = \frac{1}{LN_c} \sum_{j>i}^{N_c} \Delta L_{ij} \tag{4.1}$$

Na equação 4.1, *RCO* é a relative contact order, L é o comprimento da sequência, N_c é o número de pares de resíduos em contato (*i. e.* o número de contatos), e ΔL_{ij} é a *separação* ao longo da sequência entre os resíduos i e j em contato, ou seja, tipicamente $|i - j|$ quando os resíduos são indexados sequencialmente. A expressão foi introduzida por Plaxco, Simons e Baker em 1998[100], exibindo sucesso imediato na reprodução dos (então limitados) conjuntos de medidas experimentais de folding rate. Em trabalhos posteriores, a medida foi refinada pela introdução de nova normalização e redefinida como *absolute contact order* (*CO* na equação 4.2), exibindo performance preditiva superior em conjuntos com números maiores de medidas que incluíam proteínas de enovelamento two-state e também multistate.

$$CO = L \times RCO = \frac{1}{N_c} \sum_{j>i}^{N_c} \Delta L_{ij} \tag{4.2}$$

Desde então, (absolute) contact order se tornou essencialmente o padrão de comparação para os esforços de predição de folding rate, mas a magnitude exata da correlação é altamente sensível à composição do conjunto de dados[98]. Ademais, o autor não foi capaz de reproduzir as correlações referentes a alguns dos conjuntos de dados reportados, bem como as diferenças de performance observadas entre conjuntos de proteínas de enovelamento two-state contra multistate.

Em trabalhos posteriores, indícios importantes foram apresentados da relação entre a topologia da estrutura nativa e a topologia da estrutura do estado de transição do enovelamento[98], e a correlação entre ordem de contato e folding rate foi interpretada em termos da variação de entropia associada à reação de enovelamento[105]. Contudo, a medida de contact order tal como definida tem caráter mais de construção empírica do que de grandeza físico-química em si, e é provável que existam outras medidas com maior poder preditivo, simplicidade ou interpretabilidade. Por esta razão, buscamos desenvolver um descritor com performance superior, explorando um par de premissas simples: o fato de que em uma estrutura enovelada as distâncias inter-resíduo não são nem totalmente imprevisíveis, pois biopolímeros obedecem em média leis de escala simples, nem independentes, pois os elementos de estrutura secundária e a topologia nativa como um todo impõem vínculos rígidos entre resíduos próximos.

Aqui, apresentamos uma medida topológica simples baseada na *quantidade de informação* probabilística associada a cada contato (ou a cada *restrição*), calculada em função da verossimilhança da distância C_α-C_α observada, estimada segundo um modelo de *caminhada aleatória auto-evitante* da cadeia peptídica. Em seguida, mostramos que a medida introduzida é bem correlacionada com o logaritmo do folding rate sobre um conjunto amplo de estruturas para as quais dispomos de dados experimentais. Por fim, desenvolvemos um algoritmo simples para detectar e descartar contatos considerados redundantes do ponto de vista da topologia, e mostramos que a sua aplicação pode

aperfeiçoar a performance preditiva do descritor apresentado.

Associar uma medida probabilística de informação a cada contato inter-resíduo não é uma formulação inédita, contando com precedentes bem sucedidos na literatura. Num trabalho de 2003, Nabuurs *et al.* quantificaram a informação contida em restrições geométricas obtidas por experimentos de RMN de proteínas, calculando o aporte de informação devido à incorporação de cada nova restrição em termos do decréscimo associado da incerteza total nas posições atômicas[106]. No mesmo trabalho, Nabuurs *et al.* propõem uma estimativa da redundância entre restrições observando a variação na quantidade de informação de cada restrição em função da ordem com que a mesma é considerada, de forma que restrições mais singulares diminuem a incerteza nas posições atômicas independentemente de serem consideradas por primeiro ou por último. Crucialmente, os resultados reportados corroboram a sugestão de que as restrições mais informativas são as que impõem a aproximação de resíduos muito separados ao longo da cadeia principal, ao invés daquelas que aproximam resíduos que são vizinhos na sequência antes de o serem no espaço.

A despeito dos bons resultados, a definição de informação adotada por Nabuurs *et al.* encerra uma possibilidade evidente de sofisticação; o cálculo das incertezas antes e depois da inclusão de cada restrição não faz nenhuma hipótese sobre a distribuição de probabilidade das distâncias subjacente, desde que sejam respeitados os intervalos estabelecidos pelos valores máximos e mínimos medidos. Com isso, toda medida experimental que reduz o intervalo acessível para uma distância pelo mesmo fator traz em si a mesma quantidade de informação, independentemente dos valores de distância permitidos dentro deste novo intervalo serem de fato surpreendentes ou previsíveis. É evidente que um aumento na precisão com que se conhece uma distância constitui um aporte de informação, mas este aporte é intuitivamente menor quando a distância que passa a ser conhecida com maior precisão revela, por exemplo, que dois resíduos vizinhos na sequência estão próximos também no espaço – uma observação muito pouco surpreendente. Assim, apresentamos aqui um método alternativo para o cálculo da quantidade de informação, que contorna essa limitação. Ajustamos as distâncias interatômicas a uma distribuição de probabilidades baseada num modelo de caminhada aleatória que é função da separação entre os resíduos correspondentes ao longo da cadeia principal, atribuindo assim probabilidades menores a contatos que fecham voltas mais longas, e empregamos esta distribuição para calcular a verossimilhança de cada contato observado.

Uma caminhada aleatória auto-evitante (*self-avoiding random walk* ou SARW) é uma variante do processo de caminhada aleatória (vide [107], por exemplo) definida sobre um retículo, na qual o caminhante é proibido de retornar a posições já visitadas. A trajetória correspondente a uma realização individual e finita de uma SARW em um retículo de dimensionalidade adequada pode ser interpretada como uma conformação de um polímero linear de comprimento correspondente, uma estratégia de modelagem

empregada já por Flory em 1941[108, 109]. Exemplos são ilustrados na figura 4.3.

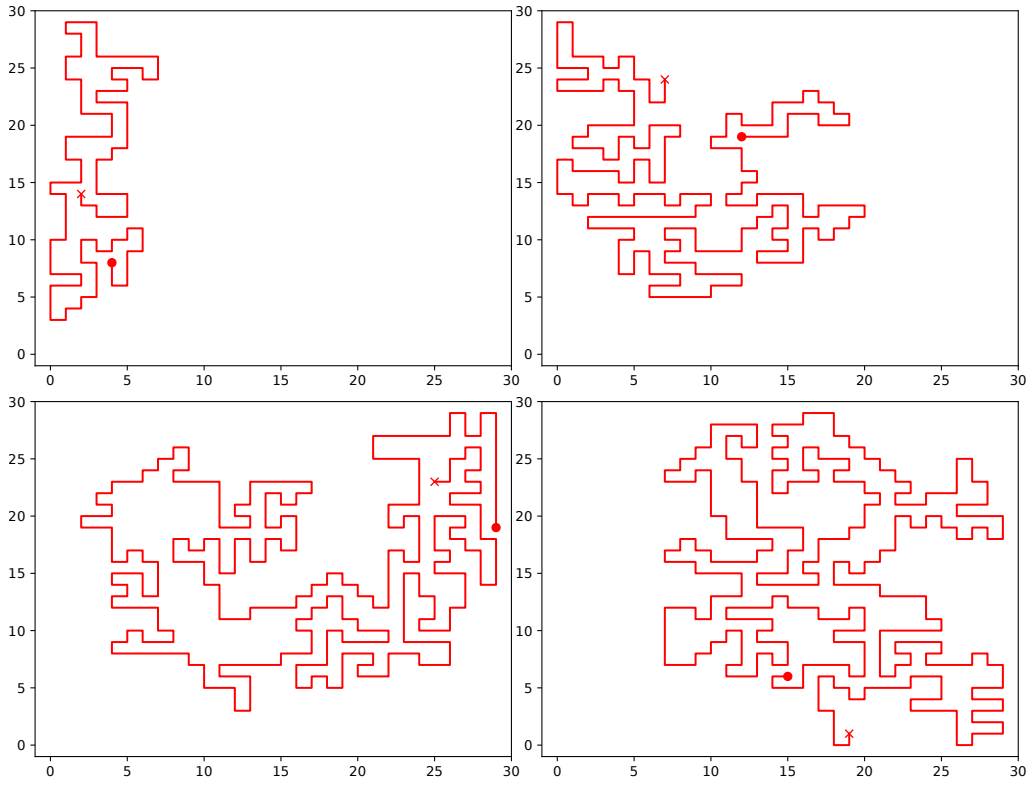


Figura 4.3. Quatro exemplos de caminhadas aleatórias auto-evitantes definidas sobre retículos bidimensionais quadrados, cujos comprimentos são, da esquerda para a direita e de cima pra baixo, de 113, 235, 343 e 352 passos. Para todas as curvas, as origens são marcadas com "×" e os términos com círculos. Definidas originalmente para modelar biopolímeros, com alguma boa vontade é possível identificar características que remetem a elementos de estrutura secundária, a contatos entre vizinhos próximos e entre monômeros distantes.

Trabalhos subsequentes como Domb *et al.*[110] e Mazur[111], ambos de 1965, abordaram as propriedades matemáticas das caminhadas auto-evitantes; em particular, a distribuição de probabilidades para a distância entre as extremidades como função de parâmetros variados. Em [111], uma expressão para essa distribuição é demonstrada como função do deslocamento médio $\langle r_N^2 \rangle$, cuja própria dependência do número N de passos percorridos deve ser fornecida independentemente. Ambos os trabalhos afirmam que essa dependência deve ter formato funcional $\langle r_N^2 \rangle^{\frac{1}{2}} = AN^{\nu}$, para valores de ν da ordem de $2\nu \approx 1{,}18$ a $1{,}25$, com alguma variação mas maiores que o valor $\nu = 1/2$ que seria característico de uma caminhada aleatória sem exclusão. De Gennes[112] e des Cloizeaux[113] oferecem desmonstrações analíticas para um expoente igual (até segunda ordem) a $\nu = 0{,}5975$, posteriormente refinado por Le Guillou e Zinn-Justin para $\nu = 0{,}588$[114], em concordância notável com o valor $\nu = 3/5$ obtido por Flory três décadas antes analisando cadeias poliméricas reais[115].

Pouco depois, Sanchez[116] demonstra que uma cadeia polimérica sujeita a interações atrativas entre monômeros e que inclui volume de exclusão exibe uma transição de fase entre um estado de cadeia expandida e um estado globular. Neste, o raio de giração, e presumivelmente a distância entre extremidades, crescem com expoente $\nu = 1/3$. Em 2004, Dima e Thirumalai mostraram que o raio de giração de proteínas no estado enovelado cresce com $N^{1/3}$ conforme esperado, porém a distância entre as extremidades *não* cresce na mesma proporção[117]. Com os resultados reportados em mente, introduzimos a expressão para $\langle r_N^2 \rangle = A\ell^{2\nu}$ na distribuição descrita por Mazur, para obter uma densidade de probabilidade para a distância entre resíduos em função do comprimento da cadeia auto-evitante que os separa. A distribuição é dada na equação 4.3, em que r é a distância observada e ℓ é a separação entre os monômeros medida em passos ao longo da cadeia.

$$f(r,\ell)dr = \frac{2{,}4343}{(A\ell^{2\nu})^{1,5}} \exp\left[-\left(0{,}8555\frac{r^2}{A\ell^{2\nu}}\right)^{1,6}\right] r^2 dr \tag{4.3}$$

Apontamos que se a equação 4.3 é usada para modelar a distância entre as extremidades de uma cadeia de aminoácidos, pode-se determinar a constante A independentemente de ν, fazendo $\ell = |i - j| = 1$ para representar um par de resíduos que são vizinhos imediatos, e ajustando A de forma que o máximo da curva ocorra em $r = 3{,}8$Å, a distância C_α-C_α ao longo de uma ligação peptídica típica. Realizamos este procedimento e obtivemos $A = 16{,}58\text{Å}^2$. Para ν, adotamos inicialmente o valor teórico $2\nu = 1{,}176$; explorações subsequentes mostraram que $2\nu = 1{,}15$ produz correlações mais altas sobre o conjunto de proteínas estudadas, mas as diferenças são de todo modo pequenas. Um exemplo da curva definida pela equação 4.3, adotando as constantes mencionadas, é apresentado na figura 4.4.

A família de distribuições obtida fornece uma expectativa aproximada para a distância euclidiana entre um par de resíduos, em termos da posição dos seus C_α, dada a sua separação em passos ao longo da cadeia principal. Seu formato traduz a noção intuitiva de que pares de resíduos separados por *loops* mais longos tendem também a se encontrar mais distantes na conformação enovelada do que pares de resíduos que são vizinhos, uma proposição que se espera verdadeira mesmo quando nenhum dos pares comparados está efetivamente em contato, dada a variação com ℓ da posição do máximo de $f(r,\ell)$.

Se a estrutura de uma dada proteína é conhecida, estas distribuições podem ser então empregadas na direção conceitualmente oposta, para avaliar a complexidade de um enovelamento contra a expectativa aleatória. Para formalizar essa noção, empregamos a medida de *quantidade de informação*[118–120] para transformar uma distância observada em uma verossimilhança da observação. A quantidade de informação de um resultado

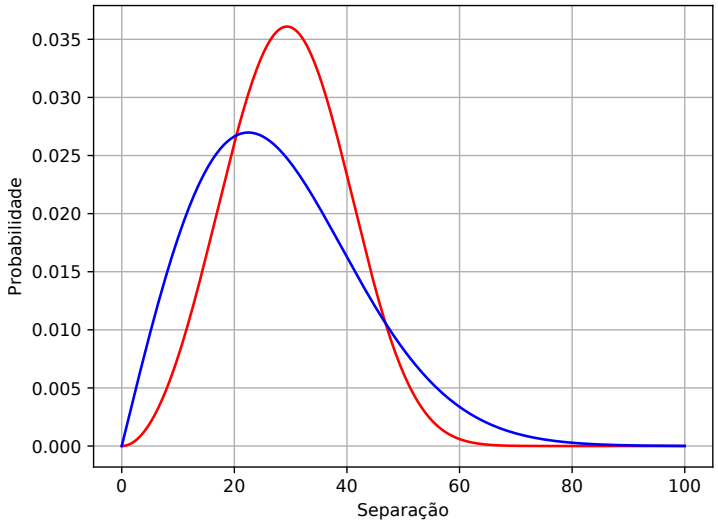


Figura 4.4. Exemplos de distribuições de probabilidades para a distância entre extremidades de caminhadas aleatórias. As duas curvas são geradas para caminhadas aleatórias de $\ell = 35$ passos, com um tamanho de passo igual a $m = 3{,}8$ unidades arbitrárias. Em azul, a distribuição de probabilidade correspondente a uma caminhada aleatória sem exclusão, dada pela distribuição normalizada $f(r,\ell) = \frac{r}{\ell m^2} e^{\frac{-r^2}{2\ell m^2}}$. Em vermelho, a distribuição dada pela equação 4.3 para as constantes análogas. Comparando as duas curvas, observa-se que o efeito do volume excluído é de aumentar a distância esperada entre as extremidades para o mesmo número de passos, ao mesmo tempo em que concentra a expectativa em torno do pico.

$X = x_i$ de uma variável aleatória *discreta* X, medida em *bits*, é dada pela equação 4.4.

$$h(x_i) = -\log_2 P(x_i) \tag{4.4}$$

A medida de quantidade de informação é uma transformação simples da probabilidade, e pela definição fica evidente que a observação de resultados menos prováveis carrega mais informação. Todavia, para calcular a quantidade de informação de uma medida experimental de distância entre os resíduos i e j, cujo valor observado está no intervalo entre d_{min} e d_{max}, é necessário reconhecer que a densidade de probabilidades dada pela equação 4.3 é contínua. Trabalhos anteriores como [106] parecem ter omitido esta sutileza ao assumir que a expressão de Shannon para a entropia de uma variável aleatória discreta pode ser generalizada para uma integral sem levar em conta a magnitude da granularidade[120], fazendo com que a entropia divirja no limite do contínuo – um descuido possivelmente cometido inclusive pelo próprio Shannon[121][5]. Aqui, ao invés de calcular o aporte de informação em termos da comparação entre incertezas prévias e posteriores, nós contornamos a questão do contínuo assumindo que as medi-

[5]Shannon reconhece que justificar rigorosamente os resultados por ele obtidos requereria "uma quantidade considerável de Teoria da Medida abstrata", mas nota que "as ocasionais liberdades tomadas com processos limite na presente análise podem ser justificadas em todos os casos de interesse prático".

das de distância tem todas a mesma precisão arbitrária, $d_{max} - d_{min} \equiv \Delta r \to 0$. Deste modo, tratamos cada observação como um resultado individual, com probabilidade associada $P_\ell(r) = \int_{d_{min}}^{d_{max}} f(r,\ell)dr \approx f(r,\ell)\Delta r$, de uma medida de distância efetivamente discreta.

Sob essas hipóteses, a quantidade de informação da distância experimental r entre os resíduos i e j é dada pela equação 4.5, na qual Δr atua como um parâmetro pequeno, arbitrário porém fixo, cujo efeito pode ser ignorado na comparação de pares de resíduos distintos.

$$h_\ell(r) = -\log_2 \left[\Delta r f(r,\ell)\right], \quad \ell = |i-j| \tag{4.5}$$

A equação 4.5 permite atribuir a cada contato, ou, no caso geral, a cada medida de distância, um juízo quantitativo sobre a verossimilhança das conformações permitidas ao trecho interposto da cadeia, cumprindo um papel similar – porém melhor embasado – ao papel da diferença ΔL_{ij} presente na definição da ordem de contato. De posse desta expressão, definimos a "Informação Topológica Média" I como a média, sobre todos os pares de resíduos, da quantidade de informação associada à distância de cada par (equação 4.6). Aqui, por simplicidade, as distâncias entre resíduos são tomadas em função das posições dos átomos C_α em todos os resultados subsequentes, e N_p representa o número de pares de resíduos.

$$I = \frac{1}{N_p}\sum_{j>i} h_\ell(d_{ij}) = -\frac{1}{N_p}\sum_{j>i} \log_2 \left[\Delta r f(d_{ij}, |i-j|)\right] \tag{4.6}$$

Investigamos, então, se era possível usar a medida de I para avaliar a complexidade da topologia de uma estrutura enovelada, e com isto obter uma estimativa do seu folding rate competitiva com aquela proveniente da ordem de contato. Os resultados são apresentados a seguir.

4.3 Correlação entre informação e taxa de enovelamento

Na seção anterior, partimos de considerações probabilísticas para introduzir a informação topológica média I, uma medida inédita e com boa interpretabilidade da complexidade da topologia de uma estrutura. Aqui, investigamos se essa medida é correlacionada com a taxa de enovelamento para proteínas em geral. Para tanto, fizemos uso de dados reunidos no banco de dados Amherst College Protein Folding Kinetics Database (ACPro)[122], tanto estruturais quanto referentes à cinética de enovelamento.

O banco de dados ACPro é uma compilação de diversos conjuntos de dados de cinética de enovelamento previamente publicados, equipado com um processo de curadoria com ênfase na verificação e precisão dos valores reportados e das condições experimentais associadas. Por ocasião da publicação destes resultados, o ACPro consiste no maior

banco de dados de cinética de enovelamento disponível; selecionamos, a partir dele, um subconjunto de 95 proteínas, descartando do conjunto original todas as proteínas com lacunas inexplicadas na indexação dos resíduos, indício de possíveis trechos faltantes da cadeia que invalidariam o cálculo da quantidade de informação, que depende dos índices dos resíduos na forma de $|i - j|$.

Para cada uma das proteínas selecionadas, obtivemos sua estrutura a partir do repositório wwPDB[4], calculamos sua informação topológica média, e comparamos com o negativo do logaritmo natural do folding rate, extraído do banco de dados. Calculamos os coeficientes de correlação de Pearson para o conjunto todo, bem como para proteínas de enovelamento two-state e multistate separadamente. Os resultados são apresentados na tabela 4.1 e na figura 4.6 (colunas centrais de cada conjunto), juntamente com intervalos de confiança de 95% calculados por *bootstrapping*. Para calcular os intervalos de confiança de cada correlação, construímos um milhão de reamostragens aleatórias de cada conjunto de estruturas e os folding rates correspondentes, calculamos os novos coeficientes de correlação, ordenamos os resultados e extraímos os valores dos 2,5% e 97,5% percentis. Incluímos, para fins de comparação, a correlação com contact order na tabela e na figura (colunas mais à esquerda de cada conjunto).

A informação topológica média aparenta ser ligeiramente mais correlacionada com o negativo do logaritmo natural do folding rate em todos os conjuntos, em comparação com a correlação com contact order. Os largos intervalos de confiança, contudo, sugerem que amostras maiores são necessárias, em particular para o subconjunto de proteínas com perfil de enovelamento multistate. Também não observamos diferenças significativas entre as correlações de contact order com folding rate para proteínas com perfil two-state, multistate, ou o conjunto todo, embora trabalhos anteriores tenham reportado que contact order se correlaciona melhor em proteínas de enovelamento multistate que as de perfil two-state[103, 104].

As correlações observadas, embora não necessariamente extraordinárias, confirmaram a expectativa de que a informação média dos contatos é uma medida válida da complexidade da estrutura nativa. Investigamos em seguida se era possível aperfeiçoar os resultados aproveitando a medida de informação topológica para identificar e descartar restrições redundantes e não-informativas, preservando ao mesmo tempo informação suficiente para caracterizar a complexidade de cada dobramento. Pra este fim, incluímos duas etapas intermediárias no cálculo da informação topológica média, definindo uma medida de "Informação Topológica Reduzida" I_r cuja performance preditiva observamos ser superior. Para calcular I_r de uma estrutura dada, identificamos e descartamos medidas de distâncias pouco informativas obedecendo aos seguintes passos:

(i) Calculamos a quantidade de informação associada a cada par de resíduos e ordenando os pares por quantidade decrescente de informação.

(ii) Descartamos todos os pares de resíduos cuja distância seja longa demais para representar um contato, considerado aqui como uma distância C_α-C_α menor ou igual a 9,5Å.

(iii) Percorremos a lista iterativamente, descartando a medida menos informativa enquanto garantimos que cada resíduo continue representado em pelo menos n medidas, com o valor $2 \leq n \leq 4$ ótimo encontrado iterativamente para cada estrutura. Repetimos este passo até que mais nenhum par possa ser descartado.

(iv) Finalmente, identificamos e descartamos pares redundantes, aplicando um critério tal que os contatos entre os pares (i,j) e (m,n) são considerados redundantes se $|i - m| + |j - n| \leq 8$. Percorremos a lista em ordem decrescente de quantidade de informação, descartando todas as medidas que são redundantes a algum contato mais informativo previamente encontrado.

No passo (iii), o valor ideal para n é $n = 4$, que corresponde ao número de medidas de distância que definem univocamente a posição de um ponto no espaço tridimensional. Contudo, para algumas estruturas é impossível exigir que todos os resíduos participem de pelo menos 4 contatos, isto é, exigir que todos os resíduos continuem sendo representados em pelo menos 4 medidas de distância após as distâncias longas terem sido descartadas. Nestes casos, adotamos o maior n possível tal que $2 \leq n \leq 4$. No passo (iv), o critério de redundância escolhido é intuitivo do ponto de vista geométrico e não requer necessariamente outra justificativa além dos bons resultados observados.

Após completar esse procedimento, a lista de contatos remanescentes contém tipicamente menos de 1% do total inicial de pares de resíduos. Um exemplo é ilustrado na figura 4.5. A informação topológica média é então calculada *sobre a lista reduzida* para obter I_r.

Calculamos I_r para o mesmo conjunto de 95 estruturas, comparando também os resultados com $-\ln(k_f)$, para o conjunto completo bem como para os subconjuntos separados por perfil de enovelamento. Os resultados são apresentados na tabela 4.1 e nas figuras 4.6 e 4.7. A distância de corte 9,5Å e o corte de separação de 8 resíduos foram determinados mediante exploração do espaço de parâmetros, e resultam na máxima correlação com o folding rate sobre o conjunto de dados, mas apontamos que a dependência da magnitude das correlações com o valor específicos de cada parâmetro é pequena, com as variações resultantes nos coeficientes de correlação sendo significativamente menores do que os intervalos de confiança de bootstrap.

Neste conjunto de dados, aproximadamente 17% de todos os pares de resíduos obedecem ao critério para serem considerados um contato. A medida de informação topológica reduzida reproduz a performance preditiva da informação topológica média, apresentando coeficientes de correlação ligeiramente maiores que os apresentados pela ordem de contato nos três recortes do conjunto, enquanto leva em consideração uma

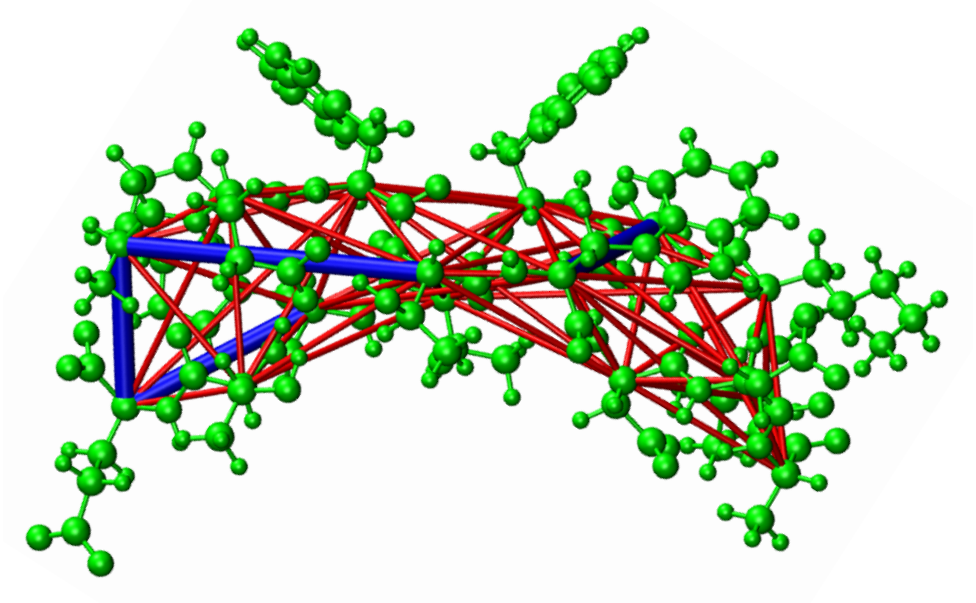


Figura 4.5. Ilustração de contatos inter-resíduos em uma proteína de 17 resíduos de comprimento. De um total de 68 contatos identificados, 64 contatos pouco informativos são representados em vermelho (tubos mais finos), e 4 contatos altamente informativos em azul (tubos mais largos), representando 5,9% do total. Figura gerada com o programa VMD[1, 2].

	Conjunto completo	*Two-state*	*Multistate*	% de pares	% de contatos
CO	0,64	0,62	0,60	17,1%	100,0%
I	0,69	0,65	0,69	17,1%	100,0%
I_r	0,68	0,67	0,68	0,7%	**4,1%**

Tabela 4.1. Resultados das correlações entre $-\ln(k_f)$ e contact order (CO), informação topológica média (I) e informação topológica reduzida (I_r) para um conjunto de 95 proteínas derivado do banco de dados ACPro. Apresentamos os coeficientes de correlação de Pearson e número médio de pares de resíduos considerados por cada medida, dado como fração de todos os pares e de todos os contatos.

média de menos de 4,1% de todos os contatos (0,7% de todos os pares de resíduos) para cada estrutura. Para uma proteína de N resíduos, com $50 \leq N \leq 400$ típico, estas frações equivalem a um número estimado entre $\sim N/8$ e $\sim N$ de pares de resíduos levados em conta, enquanto a medida de contact order leva em consideração o conjunto completo de contatos, uma média de $\sim 4N$ a $\sim 34N$ pares para proteínas no mesmo intervalo de tamanhos. Neste sentido, a medida de informação topológica reduzida representa uma melhora importante sobre resultados anteriores.

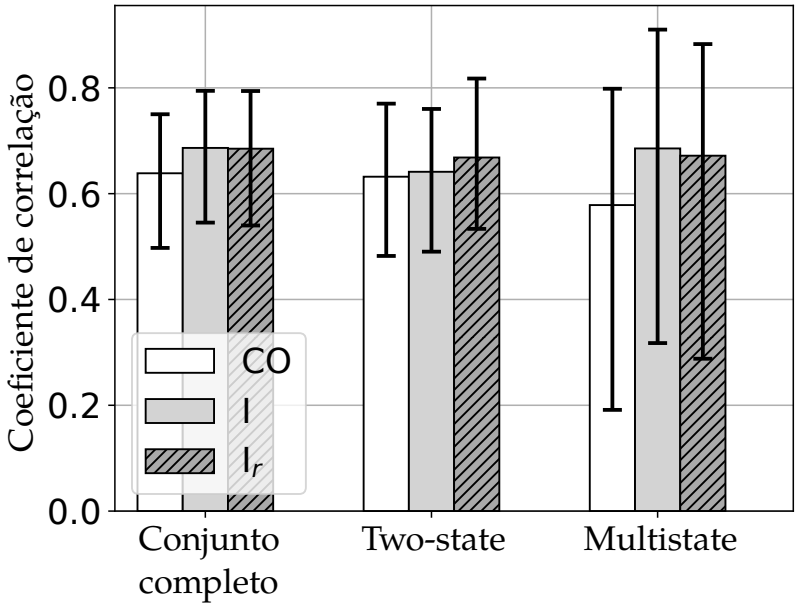


Figura 4.6. Coeficientes de correlação de Pearson entre $-\ln(k_f)$ e contact order (CO), informação topológica média (I) e informação topológica reduzida (I_r), sobre um conjunto de 95 proteínas do banco de dados ACPro. Os intervalos de confiança, no nível de 95%, foram calculados por *bootstrapping* sobre um milhão de reamostragens. A informação topológica média exibe performance preditiva comparável à de contact order em todos os casos, com coeficientes de correlação ligeiramente maiores, particularmente para proteínas com perfil de enovelamento multistate, mas com sobreposição significativa nos intervalos de confiança. A informação topológica reduzida exibe performance muito similar mas leva em consideração $\sim 96\%$ menos contatos.

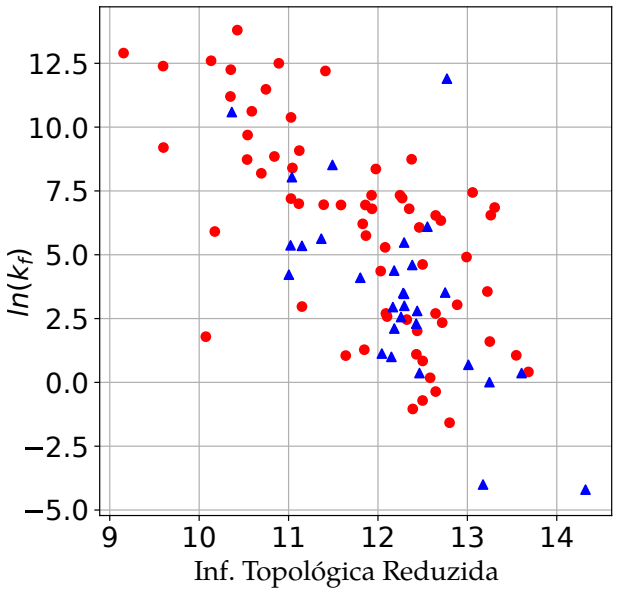


Figura 4.7. Correlação entre $\ln(k_f)$ e a informação topológica reduzida I_r, sobre um conjunto de 95 proteínas do banco de dados ACPro. Círculos representam proteínas de enovelamento two-state e triângulos representam proteínas de perfil multistate.

A título de curiosidade, também investigamos brevemente se diferenças significativas entre os valores de I e I_r para uma mesma estrutura poderiam ser atribuídas a características estruturais. A relação mais importante observada foi uma correlação ($r = 0{,}61$) entre o conteúdo de estrutura secundária e a diferença $\Delta I = I_r - I$. Observamos que ΔI aumenta conforme a fração de α-hélices decresce e a fração de folhas β aumenta. Este resultado sugere que estimativas que levam em conta *todos* os contatos podem superestimar a complexidade para proteínas compostas exclusivamente de hélices enquanto subestima a complexidade para proteínas compostas apenas por folhas.

Em conclusão, neste capítulo aplicamos ferramentas provenientes da Teoria da Informação de Shannon, associadas a um critério simples de redundância, para formalizar a diferença entre contatos informativos e não-informativos, e obtivemos com isso um descritor capaz de predizer o folding rate de uma estrutura considerando menos de 5% dos seus contatos, ressaltando o caráter redundante da maioria das interações do ponto de vista da topologia como um todo. Ressaltamos que essas observações dão embasamento também para a possível aplicação de medidas de informação para além do estudo da cinética de enovelamento em proteínas, atuando por exemplo como um mecanismo para avaliar restrições geométricas derivadas de experimentos de RMN ou de espectrometria de massas, identificando medidas críticas e possíveis erros de assinalamento. Assim, os desenvolvimentos relatados nesse capítulo também interagem com os esforços relatados no capítulo 5.

Os resultados apresentados estão publicados na revista *Bioinformatics* em [82].

Capítulo 5

Determinação de estruturas sob restrições geométricas ambíguas

Neste capítulo, relatamos os resultados da investigação que constituiu o objetivo original deste projeto: o desenvolvimento de um método resistente a erros para o cálculo de estruturas a partir de restrições geométricas derivadas de experimentos de RMN em proteínas. Em particular, buscamos implementar uma solução algorítmica que permitisse um cálculo estrutural preciso e eficiente em face de restrições ambíguas, incorretas ou incompletas, derivadas de dados experimentais com baixa relação sinal-ruído.

Embora a rigor as ambições iniciais não tenham sido atingidas, as ideias exploradas ao longo desta investigação se mostraram frutíferas do ponto de vista conceitual, e suas ramificações promoveram as origens dos trabalhos relatados nos capítulos anteriores.

Para contextualizar a metodologia proposta, apresentamos na próxima seção uma breve revisão de experimentos de RMN em proteínas.

5.1 Ressonância Magnética Nuclear em proteínas

No capítulo 1, mencionamos como a imensa maioria ($\sim 90\%$) das quase cento e cinquenta mil estruturas de proteínas depositadas no wwPDB[4] foi resolvida por meio de experimentos de difração de raios-X, com a técnica de RMN em solução, em clara desvantagem, respondendo pelos $\sim 10\%$ remanescentes[21][6].

A disparidade observada é intrigante, tendo em vista que as discussões relativas ao custo experimental de cada uma das técnicas, algo esparsas na literatura, não apontam diferença significativa entre as duas. De fato, estabelece-se um custo da ordem de \$150.000 por estrutura resolvida para casos simples, como alvos bacterianos solúveis, mas que pode chegar a quinze vezes esse valor para proteínas de membrana humanas. Alguns autores chegam inclusive a indicar que RMN é uma técnica mais barata e mais rápida para proteínas solúveis de até 25 KDa, que corresponde a aproximadamente 250 resíduos de aminoácido[124]. Na medida em que RMN é uma técnica capaz de fornecer informações dinâmicas além de estruturais, e a partir de macromoléculas em solução ao

[6]Devemos mencionar que a técnica de RMN foi em 2017 ultrapassada em número de novas estruturas por ano pela recente técnica de criomicroscopia eletrônica[123].

invés de cristais[125, 126], a sua sub-representação em termos da quantidade de estruturas resolvidas é uma questão explorada em alguma profundidade na literatura, e reflete diferenças históricas bem como dificuldades inerentes a cada aparato experimental.

Em um trabalho de 2009, Billeter *et al.*[127] argumentam que a técnica de RMN para a determinação de estruturas de proteínas incorporou a automação e a robótica em um grau muito menor que a cristalografia de proteínas, provavelmente como resultado do reconhecimento histórico da cristalografia como principal técnica para determinação de estruturas e do consequente montante de investimentos aplicados em seu desenvolvimento. De fato, Stevens afirma em [124] que o uso de robôs de cristalização diminui de dez a cem vezes o custo da produção de cristais de qualidade suficiente para difração, ao mesmo tempo em que acelera o processo. Billeter *et al.* também apontam que os dados de RMN de proteínas são tipicamente coletados próximo ao limite mínimo viável da relação sinal-ruído, incluindo sinais espúrios que dificultam tentativas de atribuição automática[127]. Os mesmos autores argumentam que, embora RMN seja uma técnica bem estabelecida para a determinação de estruturas de proteínas de até 30 KDa, não existe consenso na literatura quanto ao protocolo experimental, com os problemas sendo abordados caso-a-caso e muitas vezes dependendo de colaborações com especialistas[127]. Kay, por outro lado, afirma que existem protocolos bem estabelecidos para a determinação de estruturas de protéinas de até 50 KDa, tanto em termos dos experimentos recomendados quanto em relação à estratégia de marcação isotópica necessária[126], referindo-se principalmente a experimentos de RMN multidimensional em proteínas marcadas com ^{2}H, ^{13}C e ^{15}N[128].

Williamson & Craven desenvolvem argumentação semelhante[129], ressaltando que a automação na prática da cristalografia de proteínas se aproxima de atingir um ponto tal que os casos simples possam ser resolvidos sem a necessidade de cristalógrafos especialistas, e oferecendo um conjunto de razões pelas quais a técnica de RMN não se vê diante de panorama semelhante. Destacamos:

- A coexistência de múltiplas metodologias experimentais e a ausência de consenso quanto ao protocolo padrão, em evidente contraste com a situação da técnica de difração de raios-X, que consiste em um conjunto bem definido de passos experimentais e essencialmente uma única maneira correta de analisar os dados coletados.

- O caráter ainda incipiente dos métodos de produção automatizada de proteínas marcadas com isótopos ativos como deutério ou carbono-13, consequência da variedade de protocolos experimentais em uso, cada qual demandando estratégias diferentes de marcação – seja aleatória, completa ou mesmo dirigida a sítios específicos dentro de cada resíduo.

- A também consequente existência de múltiplos pacotes concorrentes de software para análise dos dados experimentais, muitos implementando rotinas diferentes para

alcançar objetivos semelhantes ou trabalhando com dados de entrada e saída em formato diferente.

A potencial automação dos processos de tratamento e análise dos dados experimentais, em particular, produziria benefícios tais como:

- Uma maior reprodutibilidade para estruturas calculadas a partir dos mesmos dados experimentais.
- A possibilidade de eventual identificação e correção (inclusive retroativa) de viéses sistemáticos, mesmo que sutis, no cálculo das estruturas.
- A facilitação do compartilhamento de dados entre grupos de pesquisa distintos e com o público em geral, pela adoção de um formato padrão de dados de entrada e saída específico para experimentos de RMN.
- Maior rapidez na adoção de novos métodos de análise, por meio da incorporação facilitada das rotinas correspondentes nos programas de tratamento de dados, e também na avaliação dos mesmos em função da utilização mais ampla.

O projeto se origina inserido neste cenário, visando ao aprimoramento da automação da etapa final do processo de determinação de estruturas de proteínas por RMN. Descrevemos estas etapas em seguida.

A preparação de amostras para RMN começa com a expressão e purificação de quantidades significativas da proteína de interesse. Este processo emprega técnicas de biologia molecular e é similar ao estágio inicial de produção de amostras para cristalografia, porém inclui etapas possivelmente caras de marcação isotópica. A depender do conjunto de espectros pretendido, estas etapas adicionais podem visar à incorporação de isótopos com propriedades magnéticas distintas de hidrogênio, de carbono ou nitrogênio. A marcação com deutério é empregada em praticamente todos os casos, valendo-se da razão giromagnética relativamente pequena do núcleo de ^{2}H para suprimir ressonâncias e descongestionar a região espectral do hidrogênio, bem como para eliminar mecanismos de relaxação próton-próton que tendem a alargar sinais de ressonância e promover sobreposição de picos. Contudo, quando a porcentagem de prótons que são substituídos por deutério na estrutura atinge valores próximos da totalidade, o número de interações próton-próton observáveis fica muito limitado, o que pode prejudicar a qualidade das estruturas calculadas. Existem estratégias de deuteração seletiva que se propõem a contornar essa limitação[128].

A partir das amostras marcadas, medem-se os espectros via de regra multidimensionais, em tipo e número determinados de acordo com a estratégia pretendida para o cálculo das estruturas. A análise dos dados medidos compreende três etapas a princípio sequenciais, mas que podem ser realizadas iterativamente[127]: *a*) a atribuição de cada

ressonância observada, identificada por seu valor de deslocamento quimico, a um dos átomos na estrutura, b) a análise dos picos que relacionam ressonâncias distintas, observados nos experimentos multidimensionais, para a determinação de restrições geométricas que vinculam átomos diferentes, e c) o cálculo de um conjunto de estruturas que obedecem a estas restrições. Esforços de automação existem com foco em cada uma destas etapas, bem como estratégias que abordam mais de uma ou todas simultaneamente, em caráter iterativo e retroalimentado. A identificação automatizada de picos de ressonância a partir dos dados brutos de intensidade dos espectros é, contudo, um problema técnico em aberto, cuja origem é principalmente a baixa relação sinal-ruído típica dos experimentos com proteínas[129].

Espectros bidimensionais ou multidimensionais de correlação são empregados para auxiliar tanto a atribuição dos picos de ressonância quanto a extração de restrições geométricas. Experimentos que induzem correlações pela transferência de magnetização através de ligações covalentes, tais como experimentos de TOCSY (*total correlation spectroscopy* ou "espectroscopia de correlação total"), são tipicamente utilizados para facilitar a identificação e atribuição de ressonâncias provenientes de núcleos pertencentes ao mesmo resíduo. Na figura 5.1, apresentamos um exemplo ilustrativo de um espectro de TOCSY e sua estratégia de análise. Já experimentos que detectam a transferência de magnetização através do espaço, explorando por exemplo o efeito Overhauser nuclear (NOE), são frequentemente empregados para extrair restrições geométricas de distância entre pares de prótons pertencentes a resíduos distintos e possivelmente distantes na cadeia, mas próximos no espaço. Não obstante, os espectros de correlação por NOE e sua estratégia de análise são bastante similares aos de TOCSY.

Os pacotes de análise de dados tipicamente requerem da ordem de 85% de correção e completeza na atribuição dos deslocamentos químicos para que a subsequente atribuição automática dos picos de NOE seja bem-sucedida[129]. Não obstante, via de regra a primeira passagem pela lista de picos de NOE para atribuição produz uma lista de restrições de distâncias incompleta e com inexatidões. Contudo, a depender da quantidade e porcentagem de erros das restrições obtidas, é normalmente possível a partir desta lista calcular uma ou mais estruturas candidatas, ainda que grosseiras. Alguns pacotes reportam a capacidade de calcular estruturas a partir de listas com 50% ou menos de picos reais, embora o autor tenha observado importante susceptibilidade a erros causados por restrições incorretas na sua própria metodologia. Ainda assim, tais pacotes reportam que as estruturas calculadas sob essas condições muitas vezes exibem características de estrutura secundária e classe de dobramento aproximadamente corretas, e podem ser subsequentemente utilizadas para refinar a própria atribuição dos picos de NOE [129]. O processo é repetido até a convergência segundo algum critério de parada, tipicamente a magnitude da violação residual das restrições de distância, e os resultados são reportados na forma de um ensemble de estruturas. A literatura oferece uma variedade de algorit-

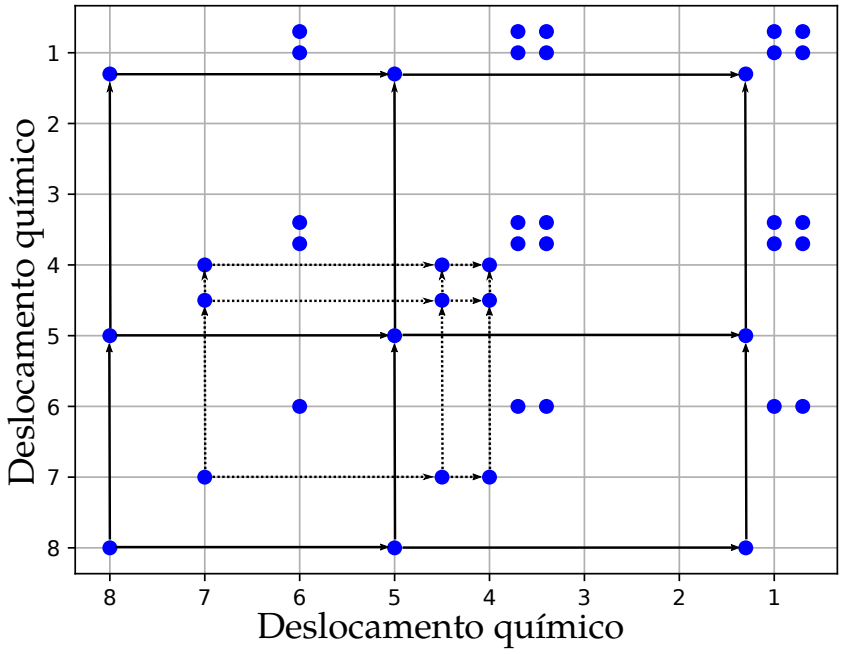


Figura 5.1. Ilustração de um espectro de TOCSY de prótons para um tripeptídeo. Os deslocamentos químicos são dados em unidades arbitrárias. Picos que compartilham linhas verticais ou horizontais são evidência da transferência de magnetização, e indicam que os núcleos correspondentes estão separados por até três ligações covalentes, via de regra por pertencerem ao mesmo resíduo de aminoácido. Seguindo as flechas é possível separar os conjuntos distintos de ressonâncias e os resíduos a que pertencem. Inspirado em uma figura de [11].

mos (e pacotes de software correspondentes), baseados em idéias diferentes, para calcular estruturas a partir de listas de restrições geométricas, e suas aplicações a dados experimentais idênticos resultam em conjuntos de estruturas com diferenças ligeiras porém sistemáticas. Para o ensemble calculado, o grau de desvio em relação à estrutura média reflete parâmetros dinâmicos e de mobilidade, mas também a precisão do resultado do cálculo[125].

Aqui, investigamos uma proposta original de algoritmo para o cálculo de estruturas a partir de restrições sujeitas a erros, cujos detalhes apresentamos na próxima seção.

5.2 Métodos de otimização para o cálculo de estruturas

O cálculo de estruturas a partir de um conjunto de restrições geométricas de alcance local não é um problema trivial do ponto de vista algorítmico, e se assemelha à montagem de um quebra-cabeças tridimensional sem uma referência da figura completa. Embora a *verificação* do grau de satisfação das restrições em uma estrutura candidata dada possa ser feita imediatamente, o espaço de estruturas possíveis é grande demais para uma exploração

exaustiva ou não-direcionada levada a cabo por um algoritmo "ingênuo", configurando essencialmente o mesmo problema mencionado na seção 4.1 que Levinthal caracterizou como o "paradoxo do dobramento"[88].

Uma abordagem tipicamente viável é a exploração do espaço conformacional guiada pela minimização de uma função energia associada a cada configuração, que guarda importantes similaridades com o processo físico-químico real. *In vitro* e *in vivo*, a depender do mecanismo de enovelamento da proteína em questão, o processo se inicia com a formação de um ou mais núcleos de enovelamento, correspondentes a interações entre pares de aminoácidos que podem ou não ser vizinhos na sequência, a partir dos quais o enovelamento da estrutura procede espontaneamente, mas a solução algorítmica do problema não fica necessariamente sujeita a tais condições. Na prática, pode-se lançar mão de movimentos drásticos como translações simultâneas de vários resíduos, rotações passando através de ângulos diedrais proibidos ou mesmo o "tunelamento" de átomos através uns dos outros, na medida em que estes recursos acelerem o decréscimo da energia.

Formalmente, a minimização da energia sem restrições é uma estratégia bem estabelecida para o refinamento de conformações correspondentes a perturbações pequenas em torno da estrutura nativa; na prática, é, no entanto, insuficiente para encontrar a estrutura nativa a partir de conformações iniciais aleatórias. A inclusão de restrições geométricas derivadas de medidas experimentais favorece o cálculo sob tais condições, impedindo movimentos que aumentariam a distância entre pares de átomos que devem ser mantidos adjacentes ou, numa implementação mais sofisticada, induzindo a aproximação dos mesmos. Para que esta indução ocorra é necessário que o algoritmo de minimização empregado seja capaz de propor movimentos que, além de diminuir a energia total, levem a conformações em que as restrições são menos violadas.

Na literatura, as estratégias variadas de cálculo existentes diferem nas soluções adotadas para o compromisso entre custo computacional e qualidade das estruturas geradas. Gardner & Kay mencionam em [128], por exemplo, um algoritmo que opera sobre um conjunto de restrições experimentais de distâncias entre pares de prótons NH-NH do backbone e assume uma função de energia puramente repulsiva, incluindo apenas termos de van der Waals, mas obtém resultados insatisfatórios. Algoritmos mais bem-sucedidos tendem a adotar potenciais completos, isto é, incluindo todos os termos de energia citados na seção 2.2, mas, principalmente, associam a cada restrição um termo próprio de energia, cujo formato mais simples é o de um potencial harmônico – análogo ao de uma ligação covalente. O formato escolhido é tal que as restrições se tornam parte da energia total, e a rotina de minimização pode valer-se do cálculo dos *gradientes* da energia para propor movimentos que levam a estruturas com baixa energia e baixo grau de violação das restrições simultaneamente. De fato, a disponibilidade dos gradientes *analíticos* dos termos do potencial interatômico e das restrições permite o emprego de rotinas de minimização muito mais eficientes do que aquelas que não calculam gradientes ou dependem

da estimativa numérica dos mesmos. O ganho de performance é tal que essencialmente inviabiliza estratégias de incorporação de restrições que não tornem acessíveis seus gradientes analíticos, limitando as maneiras com que a rotina de minimização pode lidar com restrições geométricas experimentais.

Para além dos detalhes do algoritmo, o trato com dados reais introduz um fator complicador: a eventual atribuição incorreta de picos experimentais específicos, ou a impossibilidade de atribuí-los univocamente. Descartar todas as ressonâncias de interpretação incerta é em geral ineficiente, visto que dados de RMN de proteínas são tipicamente coletados no limite mínimo prático da relação sinal-ruído[127], e erros de atribuição não são necessariamente detectáveis. Ao invés disso, tais observações podem ser preservadas na forma de restrições *ambíguas*, que podem ser incorporadas ao cálculo estrutural de duas maneiras principais: *a*) partindo de um conjunto mínimo de restrições cuja interpretação seja livre de dúvida, com o objetivo de construir um modelo grosseiro da estrutura e com este resolver as ambiguidades restantes, ou *b*) partindo de uma lista contendo todas as interpretações possíveis, visando a descartar iterativamente as restrições incorretas. Em ambos os casos, o processo via de regra não é completamente automatizado, e a intervenção do usuário é requerida principalmente para atribuir as restrições referentes a átomos das cadeias laterais[129]. Um exemplo é ilustrado na figura 5.2.

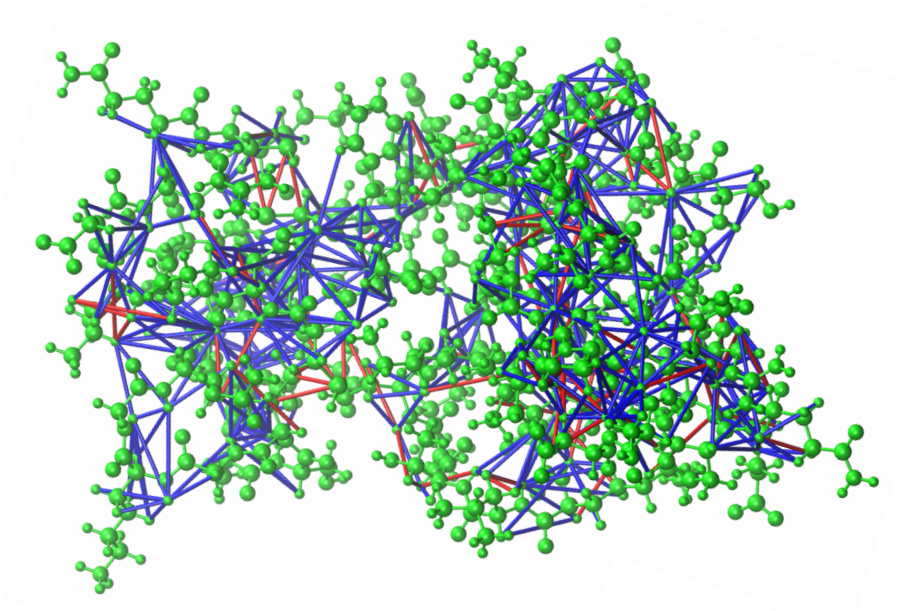


Figura 5.2. Ilustração de resultado de minimização de energia sujeita a restrições geométricas. Algumas restrições, representadas em vermelho, seguem insatisfeitas. Figura gerada com o programa VMD[1, 2].

Poderia postular-se que a simples minimização da energia sujeita a todas as restrições, corretas e incorretas, levaria ao dobramento correto da estrutura, na medida em que as restrições incorretas tendem a "puxar" em direções aleatórias, enquanto as verda-

deiras atuam em conjunto; o reforço mútuo na direção da estrutura verdadeira seria uma expressão direta do princípio da frustração mínima apresentado na seção 4.1. De fato, um dos pacotes de cálculo de estruturas de maior sucesso, o ARIA[130], emprega uma versão modificada desta ideia, onde cada pico é inicialmente traduzido como uma soma de restrições com atribuições ambíguas, ponderada pelas distâncias entre os átomos envolvidos. O ARIA incorpora também um estágio de refinamento final por dinâmica molecular em água explícita. O pacote concorrente CYANA[131] inclui um tratamento similar das atribuições ambíguas, mas adiciona uma etapa de pré-processamento inicial denominada *network anchoring*, na qual parte da ambiguidade é eliminada segundo critérios de autoconsistência do conjunto de restrições. Ambos os pacotes implementam também diversos "atalhos" heurísticos, não necessariamente justificados exceto pelos resultados obtidos, nem por vezes sequer descritos em detalhes. Ambos são simultaneamente suportados pela interface CcpNmr Analysis, importante candidata a software padrão de análise de experimentos de RMN desenvolvida pelo projeto CCPN[132], colaboração responsável pelos principais esforços de padronização e difusão em RMN de proteínas.

Aqui, investigamos a viabilidade de expressar o problema do cálculo de estruturas sujeito a restrições ambíguas como um problema de Otimização do Menor Valor Ordenado (*Low Order-Value Optimization* ou LOVO), tal qual descrito em [133, 134].

O problema da Otimização do Menor Valor Ordenado é assim definido: seja um conjunto $F = \{f_1(x), \ldots, f_r(x)\}$ de r funções concorrentes definidas sobre o domínio Ω, e seja p um inteiro tal que $p \in [1, r]$.

Se, para todo $x \in \Omega$, existe uma sequência $a_n(x)$ que é uma permutação da sequência $(1, 2, \ldots, r)$ tal que $a_n(x)$ ordena o conjunto F, fazendo com que $f_{a_1}(x) \leq f_{a_2}(x) \leq \ldots \leq f_{a_r}(x)$, então podemos definir a soma:

$$S_p(x) = \sum_{i=1}^{p} f_{a_i(x)}(x) \tag{5.1}$$

A Otimização do Menor Valor Ordenado é definida como a minimização de $S_p(x)$ sujeita a $x \in \Omega$. É possível demonstrar que, embora a soma definida na equação 5.1 não seja, no caso geral, diferenciável, $S_p(x)$ é ainda assim suscetível à aplicação de métodos de minimização baseados em gradientes[133]. A figura 5.3 ilustra o processo.

Para empregar a estratégia LOVO no contexto do cálculo de uma estrutura, incorporamos como equivalentes todas as r restrições experimentais, dentre as quais um determinado subconjunto de tamanho p corresponde a restrições verdadeiras, mas cuja identidade é desconhecida. Em cada passo da minimização, a energia de todas as restrições é calculada, e a lista de restrições é ordenada pelo valor crescente de energia. Somente as p restrições de menor energia são então consideradas naquele passo para o cálculo dos gradientes analíticos da energia, e a direção destes determina o movimento aplicado sobre a estrutura. Crucialmente, em cada passo da minimização a energia de *todas* as

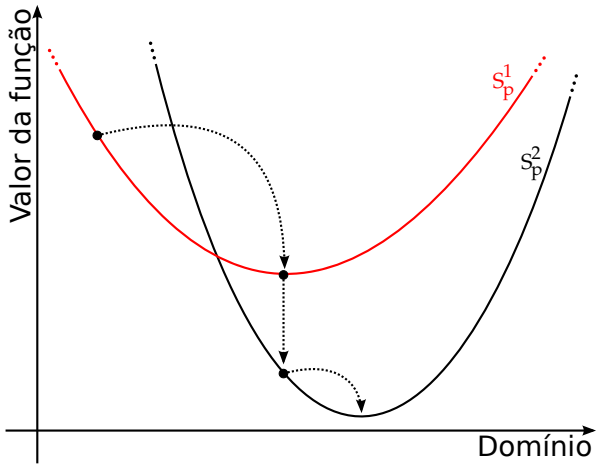


Figura 5.3. Ilustração da estratégia LOVO para a busca por um mínimo local em um conjunto de funções concorrentes. A minimização alterna movimentos na direção negativa do gradiente com a escolha da função com o menor valor na nova coordenada. Na aplicação pretendida, cada função concorrente representa a soma de um conjunto diferente de p entre r restrições geométricas. Existem $\binom{r}{p}$ tais funções. Encontrar o mínimo local implica em discernir a identidade das restrições menos violadas em casa posição, de forma a minimizar sua soma.

restrições é recalculada, e a lista é reordenada antes do cálculo dos gradientes – a soma $S_p(x)$ minimizada em cada posição é a menor entre as $\binom{r}{p}$ somas de p restrições possíveis. Presumimos que, se a estratégia é bem-sucedida, conforme a minimização se aproxima da convergência a ordem crescente das energias das restrições também para de mudar, e nos passos finais o conjunto das p restrições de menor energia corresponderá àquelas que são a interpretação verdadeira das atribuições ambíguas dos dados experimentais.

Para investigar esta hipótese, produzimos e validamos ferramentas computacionais referentes a cada etapa do processo, e relatamos os resultados na próxima seção.

5.3 Implementação da estratégia LOVO e resultados

Ao contrário das argumentações apresentadas nos capítulos 3 e 4, referentes a trabalhos fechados e publicados, nesta seção oferecemos uma narrativa de caráter mais cronológico das dificuldades enfrentadas e resolvidas na implementação e aplicação da metodologia proposta.

Inicialmente, produzimos *scripts* para a entrada e saída de dados estruturais de proteínas e de restrições geométricas, e rotinas para implementar o cálculo dos termos de energia e seus gradientes analíticos segundo o campo de força CHARMM[5], escritas em linguagem *Python* interfaceada com C para maximização da performance. Estas rotinas foram validadas por meio de comparação com as energias calculadas pelos pacotes NAMD[39] e TINKER[135] para as mesmas conformações, visando à obtenção de resul-

tados idênticos e calculados dentro de tempos computacionais comparáveis. Produzimos também rotinas de cálculo das energias das restrições e seus gradientes analíticos, em função de parâmetros ajustáveis, e validamos os resultados por meio de testes individuais e comparação com cálculos de diferença finita.

Em seguida, reunimos estruturas de tamanho pequeno e médio para atuar como modelos para os ensaios de minimização. Construimos um pequeno grupo de estruturas de teste, selecionadas a partir do repositório wwPDB[4] mediante os seguintes critérios:

- Estrutura nativa conhecida e resolvida por RMN, com os respectívos dados experimentais disponibilizados.
- Ausência de regiões não-estruturadas, tanto quanto possível.
- Ausência de ligantes, íons metálicos, modificações pós-translacionais como glicosilação ou outros fatores externos capazes de influenciar a estabilidade da estrutura nativa.
- Tamanho reduzido, porém com estrutura terciária bem definida.

Selecionamos três proteínas: um domínio Tirosina Cinase A de 107 resíduos de código 4CRP, um modelo de grampo beta mínimo com 17 resíduos de código 1LE3, e um modelo de hélice com 20 resíduos de código 2JQ0. Rapidamente ficou claro, contudo, que a estrutura de 107 resíduos impunha um custo computacional excessivamente alto para a realização de minimizações repetidas em escala, e procedemos com as análises considerando principalmente os dois peptídeos pequenos. Implementamos e executamos rotinas baseadas em algoritmos genéticos para obter também as conformações de energia mínima *global* das mesmas estruturas, embora eventualmente tais conformações tenham se mostrado mais difíceis de reproduzir durante as minimizações do que as conformações dadas como nativas, e foram abandonadas.

Para construir as configurações iniciais dos ensaios de minimização, implementamos scripts para gerar conformações iniciais aleatórias da proteína de interesse, impondo valores aleatórios para cada um de seus ângulos diedrais. Finalmente, produzimos scripts para minimizar localmente a energia de uma conformação inicial qualquer, e para organizar, executar em paralelo e analisar milhares de experimentos simultaneos de minimização de energia sob condições arbitrárias. Com estas ferramentas prontas e validadas, realizamos uma série de ensaios de minimização com o objetivo de caracterizar a performance da metodologia em função dos parâmetros utilizados.

Na discussão que segue, todos os ensaios de minimização foram feitos em relação ao peptídeo 1LE3, obedecendo ao mesmo protocolo: geramos um total de 3.000 configurações iniciais aleatórias, que são submetidas individualmente a minimização local e em seguida comparadas com a estrutura nativa conhecida para calcular sua similaridade.

Entre outras estatísticas, tabulamos a energia final de cada estrutura, a energia residual das restrições, e o TM-Score da estrutura em relação à conformação nativa (vide seção 1.3).

Contudo, antes de discutir os resultados observados, mencionamos uma dificuldade inesperada que necessitou ser contornada: a medida da qualidade e a identificação dos melhores modelos dentre os calculados, numa situação de aplicação "real" em que a estrutura alvo não é conhecida. Nos ensaios de minimização, observamos consistentemente que o valor final da energia de cada modelo produzido não apresenta correlação suficientemente alta com sua qualidade – modelos de mesma energia podem ser bastante diferentes entre si e apresentar similaridades diferentes em relação à estrutura alvo pretendida. Embora tenhamos observado que essa tendência é menos pronunciada quanto menor é a energia dos modelos em questão, outra dificuldade reside no fato de que a estrutura alvo via de regra não ocupa o mínimo global do potencial.

Essa observação não contradiz a expectativa da distribuição de Boltzmann ou a premissa do método, pois o macroestado em torno (*i. e.*, com similaridade acima de um certo limiar) da estrutura nativa pode ter um volume muito maior do que aquele em torno do mínimo global de energia potencial, resultando que o ensemble passa uma fração maior do tempo em torno da estrutura nativa apesar desta consistir em microestados individuais de probabilidade menor. Esta consideração essencialmente entrópica da superfície de energia livre se reflete nas repetidas observações da baixa correlação entre a energia potencial e a similaridade em relação à estrutura alvo.

Sem a possibilidade de empregar o valor da energia para identificar os melhores modelos, e sem um mecanismo para substituí-lo, não apenas os ensaios de minimização tem seu resultado essencialmente inutilizado, mas também a ambição mais modesta de usar os melhores modelos para identificar as restrições verdadeiras se invalida. A combinação de resultados obtidos em outros projetos dentro do grupo com uma revisão dirigida da literatura levaram à proposição de um mecanismo distinto para a medida da qualidade dos modelos calculados, e passamos a implementar um mecanismo de avaliação dos modelos baseado em *consenso*.

Medidas de consenso tem o seguinte formato geral: para cada modelo produzido, calculamos sua similaridade em relação a cada um dos outros; estas medidas são combinadas tomando, por exemplo, sua média, e o resultado é admitido como medida da qualidade do modelo. A motivação por detrás desta definição é que, desde que a amostragem seja bem sucedida o suficiente para que as regiões do espaço configuracional em torno de cada mínimo de energia, incluindo o da estrutura nativa, estejam adequadamente populadas, a quantidade de estruturas similares a cada modelo é uma boa indicação da energia livre do macroestado particular que ele representa. Neste trabalho, testamos e eventualmente adotamos uma variação da medida denominada "medida de consenso ingênuo de Davis"[136], mas a busca da expressão de consenso que produz os melhores

resultados proporcionou uma interessante investigação conceitual que não apresentaremos aqui.

Com isso, apresentamos os resultados de alguns ensaios selecionados de minimização. Num primeiro momento, realizamos ensaios livres de restrições para avaliar o comportamento das minimizações locais sujeitas apenas ao campo de força. Nestes ensaios, observamos importante dificuldade de obter bons modelos sem a inclusão de restrições geométricas (figura 5.4).

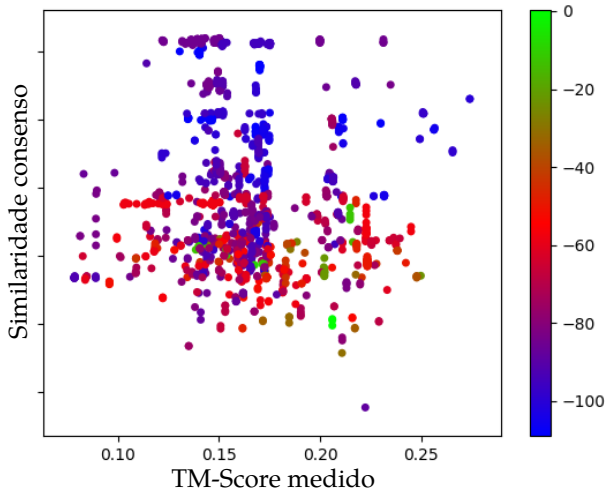


Figura 5.4. Resultado de ensaio de minimização para 3.000 conformações iniciais aleatórias do peptídeo 1LE3. Os eixos correspondem aos valores de similaridade em relação à estrutura alvo, medidos diretamente (abscissas) ou estimados por uma medida de consenso, e as cores correspondem às energias finais das estruturas. Observamos que na ausência de restrições não há amostragem próxima da estrutura nativa, conforme evidenciado pelos baixos valores de TM-Score, e as medidas de qualidade real e estimada não são correlacionadas.

Em seguida, realizamos uma série de ensaios incluindo apenas restrições verdadeiras, isto é, que correspondem a medidas de distância extraídas da estrutura nativa. Sob condições de saturação de restrições verdadeiras (figura 5.5), verificamos um rendimento modesto de estruturas aproveitáveis – com valores de TM-Score acima de 0,5 –, bem como o estabelecimento do que aparenta ser uma correlação fraca entre a qualidade das medidas e a medida de consenso.

Todavia, sob uma densidade mais realista de quatro restrições por resíduo, estruturas de qualidade aceitável ainda são obtidas mas a correlação entre as medidas de similaridade direta e de consenso é bastante prejudicada (figura 5.6).

Para contornar esta observação, investigamos a possibilidade de incluir etapas múltiplas de minimização. Determinamos que a performance dos ensaios era significativamente melhorada com a introdução de duas etapas adicionais específicas: *a*) uma etapa inicial de minimização na qual são desconsideradas as interações não-ligadas, permitindo

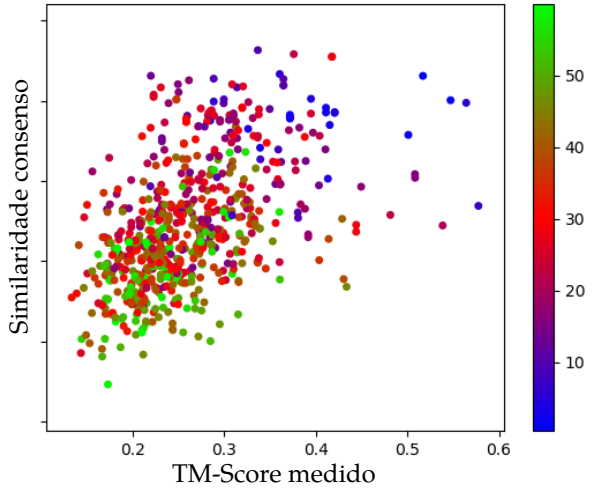


Figura 5.5. Resultado de ensaio de minimização para 3.000 conformações iniciais aleatórias do peptídeo 1LE3. Os eixos correspondem aos valores de similaridade em relação à estrutura alvo, medidos diretamente (abscissas) ou estimados por uma medida de consenso, e as cores correspondem às energias residuais *das restrições* em cada conformação. Incluindo restrições geométricas para todas as distâncias entre pares de carbonos C_α, o resultado observado é o estabelecimento de aparente correlação entre as medidas de qualidade direta e de consenso, e a obtenção de estruturas com TM-Score significativo.

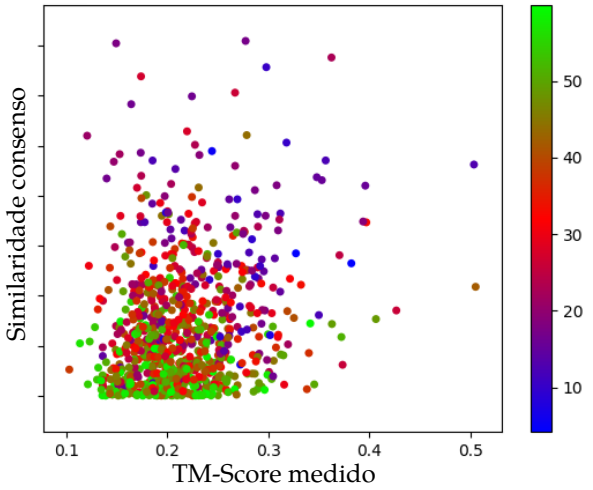


Figura 5.6. Resultado de ensaio de minimização para 3.000 conformações iniciais aleatórias do peptídeo 1LE3. Os eixos correspondem aos valores de similaridade em relação à estrutura alvo, medidos diretamente ou estimados por uma medida de consenso, e as cores correspondem às energias residuais *das restrições* em cada conformação. Considerando por volta de quatro restrições por resíduo, a correlação entre as medidas é bastante prejudicada.

movimentos drásticos na direção da satisfação das restrições geométricas, e *b*) uma etapa final de minimização na qual as restrições são desconsideradas, permitindo o relaxamento de contatos desfavoráveis sem introduzir distorções na estrutura, e com isso melhorando a correlação entre energia final e qualidade da estrutura. Novos ensaios revelaram a clara melhora da correlação após a inclusão das etapas adicionais (figura 5.7).

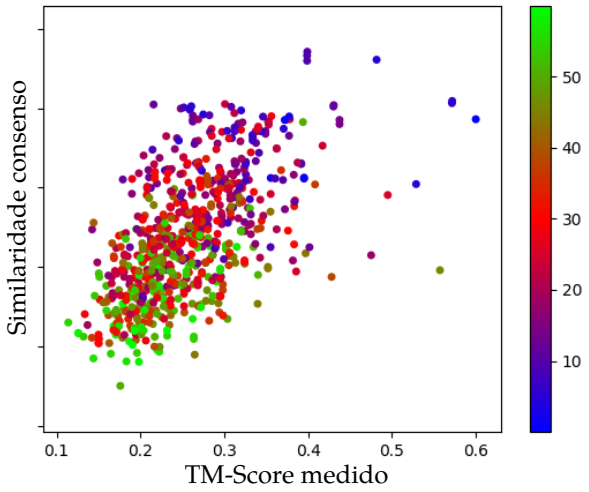


Figura 5.7. Resultado de ensaio de minimização para 3.000 conformações iniciais aleatórias do peptídeo 1LE3. Os eixos correspondem aos valores de similaridade em relação à estrutura alvo, medidos diretamente ou estimados por consenso, e as cores correspondem às energias residuais das restrições em cada conformação. Considerando por volta de quatro restrições por resíduo mas incluindo as etapas adicionais de minimização, recupera-se a correlação entre as medidas e a obtenção de estruturas com alto TM-Score.

Por fim, realizamos ensaios para determinar a melhor constante de força para o termo de energia das restrições, e observamos uma clara compensação entre o rendimento (em número de estruturas aproveitáveis) e a qualidade das estruturas produzidas (figura 5.8).

Neste ponto, passamos aos ensaios com a introdução de restrições falsas, isto é, que impõem distâncias que não são obedecidas na estrutura nativa, para avaliar a resposta das minimizações com e sem o emprego da estratégia LOVO.

Imediatamente, observamos que na presença de restrições falsas além de verdadeiras, a metodologia LOVO priorizava restrições *curtas*, isto é, restrições entre resíduos separados por poucos intermediários ao longo da cadeia, fossem elas verdadeiras ou falsas, e descartava as mais longas – provavelmente uma consequência da natureza do potencial, associada ao fato de que restrições *pouco informativas* são probabilisticamente mais fáceis de satisfazer (vide seção 4.2). Esta limitação foi contornada por meio da inclusão de uma sofisticação algorítmica similar ao conceito de *crowding*, aplicado em algoritmos genéticos para preservar a variabilidade da população ao longo de várias gerações[137]. Modificamos o algoritmo de forma a separar as restrições consideradas em conjuntos de aproximada-

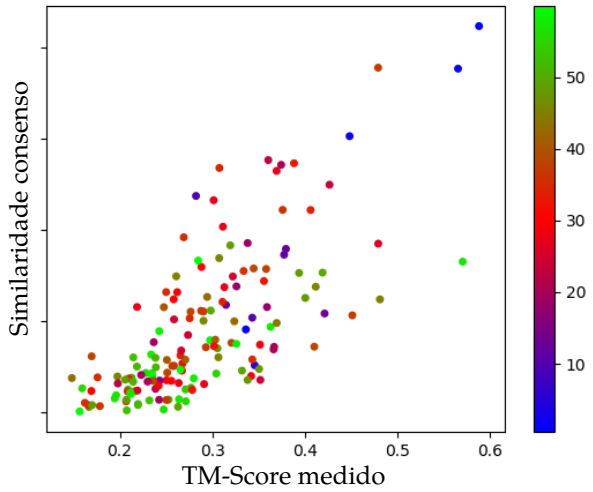


Figura 5.8. Resultado de ensaio de minimização para 3.000 conformações iniciais aleatórias do peptídeo 1LE3. Os eixos correspondem aos valores de similaridade em relação à estrutura alvo, medidos diretamente ou estimados por consenso, e as cores correspondem às energias residuais das restrições em cada conformação. Aumentando-se a constante de força das restrições em relação às das ligações covalentes, melhora-se significativamente a correlação entre as medidas, porém perde-se bastante rendimento do ensaio na medida em que muitas estruturas com quiralidades incorretas ou contatos desfavoráveis são produzidas e subsequentemente descartadas. Não calculamos TM-Score para as estruturas descartadas, e não as incluímos na figura.

mente 10 restrições cada durante a preparação da minimização. Em cada etapa de cálculo de energia, passamos a ordenar as restrições por valor crescente de energia separadamente dentro de cada conjunto, e a selecioná-las dentro de cada grupo segundo a fração inicialmente escolhida tal qual na metodologia original. Esta separação em conjuntos não altera as propriedades matemáticas desejáveis da metodologia LOVO, como a possibilidade do cálculo analítico dos gradientes da energia, e os resultados obtidos demonstraram que a modificação introduzida de fato eliminava a preferência por restrições curtas.

Contornada essa dificuldade, realizamos ensaios com a inclusão de restrições falsas além de verdadeiras, a uma proporção de 1:1, e observamos o rendimento cair a zero – a totalidade das estruturas produzidas é descartada por apresentar energias excessivamente altas ou erros na quiralidades dos carbonos C_α. Com a introdução da metodologia LOVO sob estas mesmas condições, observamos ser possível recuper algum rendimento (compare figuras 5.9 e 5.10).

Embora o rendimento assim obtido seja modesto e a correlação entre as medidas muito fraca, ao investigarmos os conjuntos de restrições escolhidas pela metodologia LOVO no final de cada minimização, observamos que estes sempre apresentam uma proporção de restrições verdadeiras maior do que os 50% da condição inicial, com uma proporção média de 72% e em muitos casos uma margem ainda mais ampla (figura 5.11).

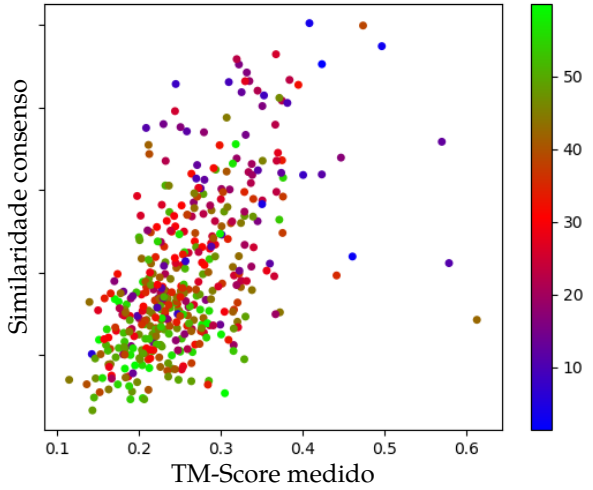


Figura 5.9. Resultado de ensaio de minimização para 3.000 conformações iniciais aleatórias do peptídeo 1LE3. Os eixos correspondem aos valores de similaridade em relação à estrutura alvo, medidos diretamente ou estimados por consenso, e as cores correspondem às energias residuais das restrições em cada conformação. Exemplo de minimização com rendimento típico, antes da introdução de restrições falsas.

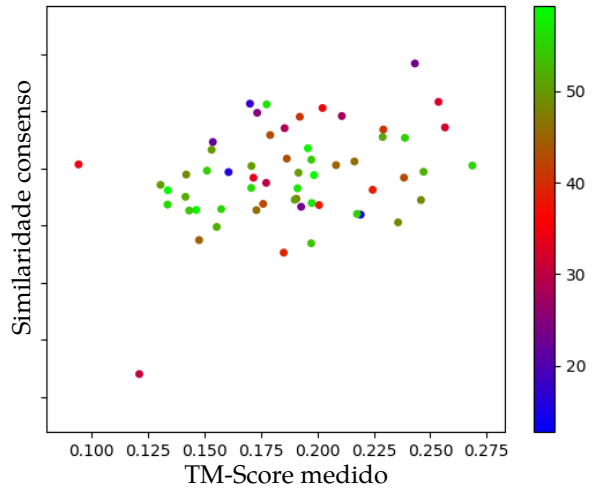


Figura 5.10. Resultado de ensaio de minimização para 3.000 conformações iniciais aleatórias do peptídeo 1LE3. Os eixos correspondem aos valores de similaridade em relação à estrutura alvo, medidos diretamente ou estimados por consenso, e as cores correspondem às energias residuais das restrições em cada conformação. A substituição de 50% das restrições por distâncias falsas leva a ensaios que não produzem nenhuma estrutura aproveitável. A subsequente introdução da metodologia LOVO nestas condições permite a recuperação de algum rendimento, aqui ilustrado. Novamente, estruturas com quiralidades incorretas ou energias excessivamente altas são descartadas, de forma que não calculamos seu TM-Score e não as incluímos na figura.

Esta observação sugere que, embora não desfrute ainda de rendimento significativo no cálculo de estruturas, a metodologia pode ser aprimorada de forma a atuar na *identificação de restrições verdadeiras* a partir de conjuntos mistos, em particular mediante aplicações sucessivas alternadas com a filtragem dos conjuntos de restrições. Deste modo, este resultado constitui talvez o maior sucesso obtido nesta investigação.

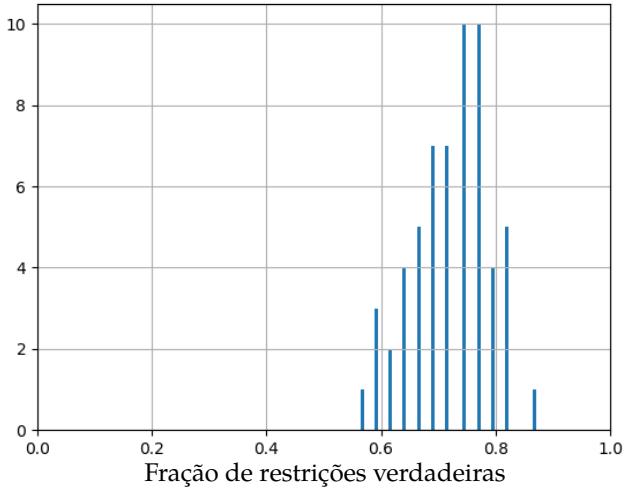


Figura 5.11. Histograma das proporções de restrições verdadeiras entre as que são consideradas para cada conformação segundo a estratégia LOVO, após um ensaio de minimização. Embora o rendimento do ensaio correspondente seja baixo, em todos os casos a metodologia seleciona um conjunto de restrições com mais restrições verdadeiras proporcionalmente do que os 50% do total fixados no início.

5.4 Conclusões

Oferecemos, então, algumas considerações finais. Os ensaios de minimização estão sujeitos a uma grande diversidade de parâmetros e condições iniciais, e, deste modo, a caracterização exaustiva da performance da metodologia proposta é um processo bastante custoso. Entretanto, os resultados observados até o momento sugerem que o cálculo da energia segundo um campo de força que inclui explicitamente todos os átomos gera uma superfície de energia potencial excessivamente rugosa, prejudicando sensivelmente a performance das minimizações locais para o cálculo de estruturas. Uma das perspectivas da continuação deste trabalho é, desse modo, a consolidação destes resultados e a subsequente adoção de um campo de força *coarse-grained*, com o objetivo de atacar estruturas maiores que somente pequenos peptídeos. Além disso, a transição para uma metodologia iterativa, baseada em sucessivas filtragens das restrições mediante ensaios de minimização com aplicação da estratégia LOVO, pode resultar no aumento gradual e significativo da proporção de restrições verdadeiras no conjunto de trabalho, e assim levar à recuperação da performance satisfatória observada nos ensaios sob restrições exclusivamente verdadei-

ras. As duas investigações devem se dar posteriormente à defesa da tese.

Finalmente, mencionamos que as ferramentas construídas durante esta investigação para os ensaios de minimização local a partir de conformações iniciais aleatórias, cuja implementação é razoavelmente eficiente do ponto de vista computacional e foi amplamente validada, foram subsequentemente empregadas pelo autor para a determinação da estrutura de um peptídeo de 18 resíduos de interesse farmacológico, como parte de uma colaboração independente que resultou num manuscrito atualmente em preparação.

Referências Bibliográficas

[1] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[2] J. Stone. An efficient library for parallel ray tracing and animation. Master's thesis, University of Missouri-Rolla, 1998.

[3] L. Censoni. Dinâmica molecular e redes complexas no estudo da difusão térmica em xilanases da família 11. Master's thesis, Universidade de São Paulo, 2013.

[4] H. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide Protein Data Bank. *Nature Structural Biology*, 10(12):980, 2003.

[5] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V.Ovochinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaeffer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: the biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, 2009.

[6] L. Martínez. Simulações de dinâmica molecular dos receptores do hormônio tireoideano, 2007.

[7] M. van Steen. *Graph theory and complex networks: an introduction.* Amsterdan: Maarten van Steen, 2010.

[8] Monty (monty@xiph.org). *Software* gPlanarity. 2018. Disponível em https://web.mit.edu/xiphmont/Public/gPlanarity.html. Acesso em 16/Nov/2018.

[9] C. Rocchini. Código-fonte distribuído sob licença Creative Commons Attribution 3.0 (https://creativecommons.org/licenses/by/3.0/). 2018. Disponível em https://www.rockini.name/math/math_images/2013/centrality.c++. Acesso em 16/Nov/2018.

[10] Domínio Público. Protein Folding Schematic. 2008. Disponível em https://commons.wikimedia.org/wiki/File:Protein_folding_schematic.png. Acesso em: 29/Nov/2018.

[11] G. S. Rule and T. K. Hitchens. *Fundamentals of Protein NMR Spectroscopy*. Springer, 2005.

[12] R. Milo. What is the total number of protein molecules per cell volume? A call to rethink some published values. *BioEssays*, 35(12):1050–1055, 2013.

[13] D. L. Nelson and M. M. Cox. *Lehninger Principles of Biochemistry*. 4th ed. New York: W. H. Freeman, 2004.

[14] T. B. Osborne. *The vegetable proteins*. London: Longmans, 1909.

[15] The Nobel Prize in Chemistry 1907. 2018. Disponível em https://www.nobelprize.org/nobel_prizes/chemistry/laureates/1907. Acesso em 12/fev/2018.

[16] The Nobel Prize in Chemistry 1946. 2018. Disponível em https://www.nobelprize.org/nobel_prizes/chemistry/laureates/1946. Acesso em 12/fev/2018.

[17] M. Brunori. 1960 *annus mirabilis*: the birth of structural biology. *Rend. Fis. Acc. Lincei*, 21(4):335–342, 2010.

[18] Y. Shi. A glimpse of structural biology through x-ray crystallography. *Cell*, 159(20):995–1014, 2014.

[19] T. L. Blundell, S. C. Harrison, R. M. Stroud, S. Yokoyama, L. E. Kay, M. G. Rossman, H. M. Berman, A. Wlodawer, E. Conti, B. Kobilka, J. M. Thornton, D. Cowburn, N. Ban, and O. Boudker. Celebrating structural biology. *Nature Structural & Molecular Biology*, 18(12):1304–1316, 2011.

[20] T. Hu, E. R. Sprague, M. Fodor, T. Stams, K. L. Clark, and S. W. Cowan-Jacob. The impact of structural biology in medicine illustrated with four case studies. *J. Mol. Med.*, 96(1):9–19, 2018.

[21] RCSB PDB - Holdings Report. 2018. Disponível em https://www.rcsb.org/pdb/statistics/holdings.do. Acesso em 16/fev/2018.

[22] G. Wei, W. Xi, R. Nussinov, and B. Ma. Protein ensembles: How does nature harness thermodynamic fluctuations for life? The diverse functional roles of conformational ensembles in the cell. *Chemical Reviews*, 116(11):6516–6551, 2016.

[23] K. Henzler-Wildman and D. Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, 2007.

[24] J. N. Onuchic and P. G. Wolynes. Theory of protein folding. *Current Opinion in Structural Biology*, 14(1):70–75, 2004.

[25] G. M. Alter. Comparison of solution and crystalline state protein structures. *The Journal of Biological Chemistry*, 258(24):14966–14973, 1983.

[26] B. Rupp. *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. Garland Science, 2009.

[27] R. B. Best, K. Lindorff-Larsen, M. A. DePristo, and M. Vendruscolo. Relation between native ensembles and experimental structures of proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 103(29):10901–10906, 2006.

[28] K. Sikic, S. Tomic, and O. Carugo. Systematic comparison of crystal and NMR protein structures deposited in the protein data bank. *The Open Biochemistry Journal*, 4:83–95, 2010.

[29] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976.

[30] J. Xu and Y. Zhang. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, 26(10):889–895, 2010.

[31] Y. Zhang and J. Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710, 2004.

[32] I. Kufareva and R. Abagyan. *Methods of Protein Structure Comparison*, pages 231–257. Humana Press, New York, 2012.

[33] J. von Plato. Boltzmann's ergodic hypothesis. *Archive for History of Exact Sciences*, 42(1):71–89, 1991.

[34] C. R. de Oliveira and T. Werlang. Ergodic hypothesis in classical statistical mechanics. *Rev. Bras. Ensino Fís.*, 29(2):189–201, 2007.

[35] F. Reif. *Fundamentals of statistical and thermal physics*. McGRAW-HILL BOOK COMPANY, 1965.

[36] E. T. Jaynes. Gibbs vs Boltzmann Entropies. *Am. J. Phys.*, 33(5):391–398, 1965.

[37] S. Riniker. Fixed-charge atomistic force fields for molecular dynamics simulations in the condensed phase: An overview. *J. Chem. Inf. Model.*, 58(3):565–578, 2018.

[38] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, 2003.

[39] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[40] P. H. Hünenberger. *Thermostat Algorithms for Molecular Dynamics Simulations*, volume 173, pages 105–149. Springer, 2005.

[41] N. Ota and D. A. Agard. Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. *Journal of Molecular Biology*, 351(2):345–354, 2005.

[42] L. Martínez, A. C. M. Figueira, P. Webb, I. Polikarpov, and M. S. Skaf. Mapping the intramolecular vibrational energy flow in proteins reveals functionally important residues. *The Journal of Physical Chemistry Letters*, 2(6):2073–2078, 2011.

[43] L. Censoni, H. S. Muniz, and L. Martínez. A network model predicts the intensity of residue-protein thermal coupling. *Bioinformatics*, 33(14):2106–2113, 2017.

[44] D. Leitner. Energy flow in proteins. *Annual Review of Physical Chemistry*, 59(1):233–259, 2008.

[45] K. Moritsugu, O. Miyashita, and A. Kidera. Vibrational energy transfer in a protein molecule. *Physical Review Letters*, 85(18):3970–3973, 2000.

[46] K. A. Peterson, C. W. Rella, J. R. Engholm, and H. A. Schwettman. Ultrafast vibrational dynamics of the myoglobin amide i band. *J. Phys. Chem. B*, 103(3):557–561, 1999.

[47] S. Lampa-Pastirk and W. F. Beck. Intramolecular vibrational preparation of the unfolding transition state of Zn^{ii}-substituted cytochrome *c*. *J. Phys. Chem. B*, 110(46):22971–22974, 2006.

[48] L. Bleicher, E. T. Prates, T. C. F. Gomes, R. L. Silveira, A. S. Nascimento, A. L. Rojas, A. Golubev, L. Martínez, M. S. Skaf, and I. Polikarpov. Molecular basis of the thermostability and thermophilicity of laminarinases: X-ray structure of the hyperthermostable laminarinase from *Rhodothermus marinus* and molecular dynamics simulations. *J. Phys. Chem. B*, 115(24):7940–7949, 2011.

[49] A. A. S. T. Ribeiro and V. Ortiz. A chemical perspective on allostery. *Chemical Reviews*, 116(11):6488–6502, 2016.

[50] R. A. Laskowski, F. Gerick, and J. M. Thornton. The structural basis of allosteric regulation in proteins. *FEBS Letters*, 583(11):1692–1698, 2009.

[51] S. Tzeng and C. G. Kalodimos. Protein dynamics and allostery: an NMR view. *Current Opinion in Structural Biology*, 21(1):62–67, 2011.

[52] H. N. Motlagh, J. O. Wrabl, J. Li, and V. J. Hilser. The ensemble nature of allostery. *Nature*, 508(7496):331–339, 2014.

[53] M. Scian, M. Acchione, M. Li, and W. M. Atkins. Reaction dynamics of ATP hydrolysis catalyzed by p-glycoprotein. *Biochemistry*, 53(6):991–100, 2014.

[54] J. Liang and K. A. Dill. Are proteins well-packed? *Biophysics Journal*, 81(10):751–766, 2001.

[55] M. E. Rhodes and M. J. Blunt. Advective transport in percolation clusters. *Phys. Rev. E*, 75(1):011124, 2007.

[56] S. Brüschweiler, P. Schanda, K. Kloiber, B. Brutscher, G. Kontaxis, R. Konrat, and M. Tollinger. Direct observation of the dynamic process underlying allosteric signal transmission. *J. Am. Chem. Soc.*, 131(8):3063–3068, 2009.

[57] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[58] S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[59] S. B. Johnson. *Emergence: The Connected Lives of Ants, Brains, Cities, and Software*. New York: Scribner, 2002.

[60] A. Krishnan, J. P. Zbilut, M. Tomita, and A. Giuliani. Proteins as networks: usefulness of graph theory in protein science. *Current Protein and Peptide Science*, 9(1):28–38, 2008.

[61] C. Böde, I. A. Kovács, M. S. Szalayb, R. Palotaib, T. Korcsmárosb, and P. Csermely. Network analysis of protein dynamics. *FEBS Letters*, 581(15):2776–2782, 2007.

[62] P. Csermely, K. S. Sandhu, E. Hazai, Z. Hoksza, H. J. M. Kiss., F. Miozzo, D. V. Veres, F. Piazza, and R. Nussinov. Disordered proteins and network disorder in network descriptions of protein structure, dynamics and function. Hypotheses and a comprehensive review. *Current Protein and Peptide Science*, 13(1):19–33, 2012.

[63] S. M. Patra and S. Vishveshwara. Backbone cluster identification in proteins by a graph theoretical method. *Biophysical Chemistry*, 84(1):13–25, 2000.

[64] P. J. Artymiuk, A. R. Poirrette, H. M. Grindley, D. W. Rice, and P. Willett. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *Journal of Molecular Biology*, 243(2):327–344, 1994.

[65] J. Huan, D. Bandyopadhyay, W. Wang, J. Snoeyink, J. Prins, and A. Tropsha. Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. *Journal of Computational Biology*, 12(6):657–671, 2005.

[66] N. Kannan and S. Vishveshwara. Identification of side-chain clusters in protein structures by a graph spectral method. *Journal of Molecular Biology*, 292(2):441–464, 1999.

[67] N. V. Dokholyan, L. Li, F. Ding, and E. I. Shakhnovich. Topological determinants of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 99(13):8637–8641, 2002.

[68] L.H. Greene and V.A. Higman. Uncovering network systems within protein structures. *Journal of Molecular Biology*, 334(4):781–791, 2003.

[69] L. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: a survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

[70] J. L. Moreno. *Who Shall Survive?* Beacon: Beacon House, 1934.

[71] A. Bavelas. A mathematical model for group structures. *Human Organization*, 7(3):16–30, 1948.

[72] L. C. Freeman. Centrality in social networks: conceptual clarification. *Social Networks*, 1(3):215–239, 1978.

[73] M. Benzi and C. Klymko. On the limiting behavior of parameter-dependent network centrality measures. *SIAM J. ANAL. APPL.*, 36(2):686–706, 2015.

[74] M. Vendruscolo, N. V. Dokholyan, E. Paci, and M. Karplus. Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E*, 65(6):061910, 2002.

[75] M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus. Three key residues form a critical contact network in a protein folding transition state. *Nature*, 409(6820):641–645, 2001.

[76] A. del Sol and P. O'Meara. Small-world network approach to identify key residues in protein-protein interaction. *Proteins: structure, function, and bioinformatics*, 58(3):672–682, 2005.

[77] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, and S. Pietrokovski. Network analysis of protein structures identifies functional residues. *Journal of Molecular Biology*, 344(4):1135–1146, 2004.

[78] H. Muniz. Estudo computacional da difusão térmica em proteínas termoestáveis. Master's thesis, Universidade de São Paulo, São Carlos, 2013.

[79] K. Z. Szalay and P. Csermely. Perturbation centrality and turbine: A novel centrality measure obtained using a versatile network dynamics tool. *PLoS ONE*, 8(10):e78059, 2013.

[80] Y. Moreno and A. F. Pacheco. Synchronization of Kuramoto oscillators in scale-free networks. *Europhys. Lett.*, 68(4):603–609, 2004.

[81] M. Benzi and C. Klymko. Total communicability as a centrality measure. *Journal of Complex Networks*, 1(2):124–149, 2013.

[82] L. Censoni and L. Martínez. Prediction of kinetics of protein folding with non-redundant contact information. *Bioinformatics*, 34(23):4034–4038, 2018.

[83] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[84] C. B. Anfinsen and E. Haber. Studies on the reduction and re-formation of protein disulfide bonds. *J. Biol. Chem.*, 236(5):1361–1363, 1961.

[85] V. S. Pande, A. Y. Grosberg, and T. Tanaka. Nonrandomness in protein sequences: evidence for a physically driven stage of evolution? *Proc. Natl. Acad. Sci. U. S. A.*, 91(26):12972–12975, 1994.

[86] J. L. England and G. Haran. Role of solvation effects in protein denaturation: From thermodynamics to single molecules and back. *Annu. Rev. Phys. Chem.*, 62:257–277, 2011.

[87] C. Camilloni, D. Bonetti, A. Morrone, R. Giri, C. M. Dobson, M. Brunori, S. Gianni, and M. Vendruscolo. Towards a structural biology of the hydrophobic effect in protein folding. *Scientific Reports*, 6(28285):1–9, 2016.

[88] C. Levinthal. How to fold graciously. *Mössbaun Spectroscopy in Biological System Proceedings*, 67(41):22–24, 1969.

[89] M. Karplus and D. L. Weaver. Protein folding dynamics: The diffusion-collision model and experimental data. *Protein Science*, 3(4):650–668, 1994.

[90] J. Jacob, B. Krantz, R. S. Dothager, P. Thiyagarajan, and T. R. Sosnick. Early collapse is not an obligate step in protein folding. *J. Mol. Biol.*, 338(2):369–382, 2004.

[91] V. Daggett and A. R. Fersht. Is there a unifying mechanism for protein folding? *TRENDS in Biochemical Sciences*, 28(1):18–25, 2003.

[92] O. V. Galzitskaya, S. O. Garbuzynskiy, and A. V. Finkelstein. Theoretical study of protein folding: outlining folding nuclei and estimation of protein folding rates. *J. Phys.: Condens. Matter*, 17(18):S1539–S1551, 2005.

[93] S. E. Jackson and A. R. Fersht. Folding of chymotrypsin inhibitor 2. 1. evidence for a two-state transition. *Biochemistry*, 30(43):10428–10435, 1991.

[94] S. E. Jackson. How do small single-domain proteins fold? *Folding & Design*, 3(4):R81–R91, 1998.

[95] H. Kaya, Z. Liu, and H. S. Chan. Chevron behavior and isostable enthalpic barriers in protein folding: Successes and limitations of simple gō-like modeling. *Biophysical Journal*, 89(1):520–535, 2005.

[96] H. Kaya and H. S. Chan. Origins of chevron rollovers in non-two-state protein folding kinetics. *Physical Review Letters*, 90(25):258104, 2003.

[97] P. Kukic, Y. Pustovalova, C. Camilloni, S. Gianni, D. M. Korzhnev, and M. Vendruscolo. Structural characterization of the early events in the nucleation-condensation mechanism in a protein folding process. *J. Am. Chem. Soc.*, 139(20):6899–6910, 2017.

[98] E. Paci, K. Lindorff-Larsen, C. M. Dobson, M. Karplus, and M. Vendruscolo. Transition state contact orders correlate with protein folding rates. *J. Mol. Biol.*, 352(3):495–500, 2005.

[99] O. V. Galzitskaya, S. O. Garbuzynskiy, D. N. Ivankov, and A. V. Finkelstein. Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *PROTEINS: Structure, Function, and Genetics*, 51(2):162–166, 2003.

[100] K. W. Plaxco, K. T. Simons, and D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, 277(4):985–994, 1998.

[101] K. W. Plaxco, K. T. Simons, I. Ruczinski, and D. Baker. Topology, stability, sequence, and length: Defining the determinants of two-state protein folding kinetics. *Biochemistry*, 39(37):11177–11183, 2000.

[102] V. Grantcharova, E. J. Alm, D. Baker, and A. L. Horwich. Mechanisms of protein folding. *Curr. Opin. Struct. Biol.*, 11(1):70–82, 2001.

[103] D. N. Ivankov, S. O. Garbuzynskiy, E. Alm, K. W. Plaxco, D. Baker, and A. V. Finkelstein. Contact order revisited: Influence of protein size on the folding rate. *Protein Science*, 12(9):2057–2062, 2003.

[104] A. S. Wagaman and S. S. Jaswal. Capturing protein folding-relevant topology via absolute contact order variants. *Journal of Theoretical and Computational Chemistry*, 13(1):1450005, 2014.

[105] A. V Finkelstein and A. Y. Badretdinov. Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Folding & Design*, 2(2):115–121, 1997.

[106] S. B. Nabuurs, C. A. E. M. Spronk, E. Krieger, H. Maassen, G. Vriend, and G. W. Vuister. Quantitative evaluation of experimental NMR restraints. *J. Am. Chem. Soc.*, 125(39):12026–12034, 2003.

[107] K. Pearson. The problem of the random walk. *Nature*, 72(1865):294, 1905.

[108] P. J. Flory. Thermodynamics of high polymer solutions. *J. Chem. Phys.*, 9(8):660–661, 1941.

[109] P. J. Flory. Thermodynamics of high polymer solutions. *J. Chem. Phys.*, 10(1):51–61, 1942.

[110] C. Domb, J. Gillis, and G. Wilmers. On the shape and configuration of polymer molecules. *Proc. Phys. Soc.*, 85(4):625–645, 1965.

[111] J. Mazur. Distribution function of the end-to-end distances of linear polymers with excluded volume effects. *Journal of Research of the National Bureau of Standards - A. Physics and Chemistry*, 69A(4):355–363, 1965.

[112] P. G. de Gennes. Exponents for the excluded volume problem as derived by the wilson method. *Physics Letters*, 38A(5):339–340, 1972.

[113] J. des Cloizeaux. Lagrangian theory for a self-avoiding random chain. *Physical Review A*, 10(5):1665–1669, 1974.

[114] J. C. Le Guillou and J. Zinn-Justin. Critical exponents from field theory. *Physical Review B*, 21(9):3976–3998, 1980.

[115] P. J. Flory. The configuration of real polymer chains. *J. Chem. Phys.*, 17(3):303–310, 1949.

[116] I. C. Sanchez. Phase transition behavior of the isolated polymer chain. *Macromolecules*, 12(5):980–988, 1979.

[117] R. I. Dima and D. Thirumalai. Asymmetry in the shapes of folded and denatured states of proteins. *J. Phys. Chem.*, 108(21):6564–6570, 2004.

[118] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[119] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(4):623–656, 1948.

[120] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[121] E. T. Jaynes. *Information Theory and Statistical Mechanics*, pages 181–218. Benjamin, New York, 1963.

[122] A. S. Wagaman, A. C., I. Brand-Thomas, B. Dash, and S. S. Jaswal. A comprehensive database of verified experimental data on protein folding kinetics. *Protein Science*, 23(12):1808–1812, 2104.

[123] S. C. Shoemaker and N. Ando. X-rays in the cryo-electron microscopy era: Structural biology's dynamic future. *Biochemistry*, 57(3):277–285, 2018.

[124] R. C. Stevens. The cost and value of three-dimensional protein structure. *Drug Discovery World*, pages 35–48, 2003.

[125] K. Wüthrich. Protein structure determination in solution by Nuclear Magnetic Resonance spectroscopy. *Science*, 243(4887):45–50, 1989.

[126] L. E. Kay. NMR studies of protein structure and dynamics. *Journal of Magnetic Resonance*, 173(2):193–207, 2005.

[127] M. Billeter, G. Wagner, and K. Wüthrich. Solution NMR structure determination of proteins revisited. *Journal of Biomolecular NMR*, 42(3):155–158, 2008.

[128] K. H. Gardner and L. E. Kay. THE USE OF 2H, 13C, 15N MULTIDIMENSIONAL NMR TO STUDY THE STRUCTURE AND DYNAMICS OF PROTEINS. *Annual Review of Biophysics and Biomolecular Structure*, 27:357–406, 1998.

[129] M. P. Williamson and C. J. Craven. Automated protein structure calculation from NMR data. *Journal of Biomolecular NMR*, 43(3):131–143, 2009.

[130] W. Rieping, M. Habeck, B. Bardiaux, A. Bernard, T. E. Malliavin, and M. Nilges. ARIA2: Automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics*, 23(3):381–382, 2007.

[131] P. Güntert and L. Buchner. Combined automated NOE assignment and structure calculation with CYANA. *Journal of Biomolecular NMR*, 62(4):453–471, 2015.

[132] S. P. Skinner, B. T. Goult, R. H. Fogh, W. Boucher, T. J. Stevens, E. D. Laue, and G. W. Vuister. Structure calculation, refinement and validation using CcpNmr Analysis. *Acta Crystallographica Section D Biological Crystallography*, 71(1):154–161, 2015.

[133] R. Andreani, J. M. Martínez, L. Martínez, and F. Yano. Low Order-Value Optimization and applications. *Journal of Global Optimization*, 43(1):1–22, 2009.

[134] L. Martínez, R. Andreani, and J. M. Martínez. Convergent algorithms for protein structural alignment. *BMC Bioinformatics*, 8(1):306–320, 2007.

[135] J. W. Ponder and F. M. Richards. An efficient newton-like method for molecular mechanics energy minimization of large molecules. *J. Comput. Chem.*, 8(7):1016–1024, 1987.

[136] A. Kryshtafovych, B. Monastyrskyy, K. Fidelis, T. Schwede, and A. Tramontano. Assessment of model accuracy estimations in CASP12. *Proteins: Structure, Function, and Bioinformatics*, 86(S1):345–360, 2018.

[137] K. A. de Jong. An analysis of the behavior of a class of genetic adaptive systems, 1975.